PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-242434
(43)Date of publication of application : 08.09.2000

(51)Int.Cl. G06F 3/06

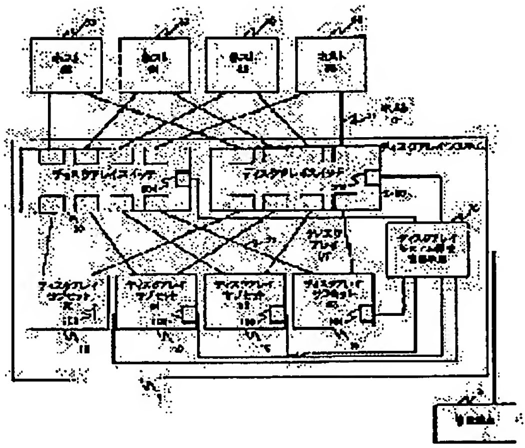
(21)Application number : 11-344260
(22)Date of filing : 03.12.1999

(71)Applicant : HITACHI LTD
(72)Inventor : MATSUNAMI NAOTO
OEDA TAKASHI
YAMAMOTO AKIRA
AJIMATSU YASUYUKI
SATO MASAHIKO

(30)Priority
Priority number : 10364079 Priority date : 22.12.1998 Priority country : JP

(54) STORAGE DEVICE SYSTEM

(57)Abstract:
PROBLEM TO BE SOLVED: To construct a storage device system corresponding to the scale or request of a computer system so that the extension of a storage device system and improvement in reliability in the future are easily realized.
SOLUTION: This system 1 has a plurality of subsets 10 having a storage device for holding data and a controller for controlling the storage device and switch devices 20 arranged between the subsets 10 and a host 30. Each switch device 20 has a managing table for holding management information for managing the configuration of the storage device system 1. According to the management information, address information contained in frame information outputted by the host 30 is translated and the frame information is distributed to the subsets 10.



LEGAL STATUS

[Date of request for examination] 24.12.2003

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

CLAIMS

[Claim(s)]
[Claim 1] Two or more store subsystems which have the store which has a storage holding data, and the control device which controls this store, The configuration management table which was connected to the calculating machine which uses the data held at said two or more store subsystems, and stored store structure-of-a-system information, The 1st interface node which has a frame conversion means to change said frame based on the configuration information which answered the frame sent from said calculating machine, analyzed this frame, and was held at said configuration management table, Two or more 2nd interface nodes by which each was connected to any one of said the store subsystems, The storage system characterized by having a transfer means for said 1st interface node and said two or more 2nd interface nodes to be connected, and to transmit said frame between said 1st interface node and said two or more 2nd interface nodes.
[Claim 2] It is the storage system according to claim 1 which said 1st interface node has a packet generation means to add and output the node address information on said 2nd interface node to said frame, and is characterized by said transfer means transmitting said frame based on said node address information between said 1st interface node and said two or more 2nd interface nodes.
[Claim 3] It is the storage system according to claim 1 which said frame has a header unit holding the identifier which specifies the source and the destination, and the data stereo section holding the stereo data transmitted, and is characterized by said conversion means changing the identifier of the destination held at said header unit based on said configuration information.
[Claim 4] Said conversion means is a storage process defined system according to claim 3 characterized by changing said 1st logical address information into the 2nd logical address managed within the store subsystem used as the destination of this frame based on said configuration information held at said configuration management table including the 1st logical address information that said frame is recognized by said data stereo section by said computer.
[Claim 5] Said storage system is a storage system according to claim 1 characterized by connecting with said transfer means, receiving the input of the configuration information which defines the storage structure of a system from an operator, answering this input further, and having the management processor which sets said configuration information as said configuration management table of each node.
[Claim 6] Said configuration information is a storage process defined system according to claim 5 characterized by including the information which restricts access to said two or more store subsystems from said calculating machine.
[Claim 7] Said 1st interface node is a storage process defined system according to claim 2 characterized by to add node address information which is different on each frame, and to transmit to said transfer means so that the light command frame which directs the writing of the data transmitted from said calculating machine may be answered, those duplicates may be generated about the data frame following this light command frame and it and the data frame following said light command frame and it may be sent to at least two store subsystems.
[Claim 8] Said 1st interface node is a store subsystem according to claim 7 characterized by adding node address information which is different in each lead command frame, and transmitting to said transfer means so that the lead command frame which directs the lead of the data transmitted from said calculating machine may be answered, the duplicate of this lead command frame may be generated and said lead command frame may be sent to said at least two store subsystems.
[Claim 9] Said 1st interface node is a storage process defined system according to claim 8 characterized by receiving the data frame which answers said lead command frame and is transmitted from said at least two store subsystems, choosing one of these, and transmitting to said computer.
[Claim 10] Said 1st interface node is a store subsystem according to claim 7 characterized by answering the lead command frame which directs the lead of the data transmitted from said calculating machine,

adding the node address information on the 2nd interface node linked to the store subsystem of 1 beforehand defined between said at least two store subsystems to said lead command frame, and transmitting to said transfer means.

[Claim 11] Two or more store subsystems which have the store which has a storage holding data, and the control device which controls this store, The configuration management table which is switching equipment connected between the calculating machines using the data stored in said store, was connected to said calculating machine, and stored store structure-of-a-system information, A conversion means to change said frame based on said configuration information which answered the frame sent from said calculating machine, analyzed this frame, and was held at said configuration management table, Two or more 2nd interface nodes by which each was connected to either of said store subsystems, Switching equipment by which it is having-transfer means for said 1st interface node and said two or more 2nd interface nodes to be connected, and to transmit said frame between said 1st interface node and said two or more 2nd interface nodes characterized.

[Claim 12] It is switching equipment according to claim 11 which has a packet generation means by which said 1st interface node adds and outputs the node address information on said 2nd interface node to said frame, and is characterized by said transfer means transmitting said frame based on said node address information between said 1st interface node and said two or more 2nd interface nodes.

[Claim 13] It is switching equipment according to claim 11 which said frame has a header unit holding the identifier which specifies the source and the destination, and the data stereo section holding the stereo data transmitted, and is characterized by said conversion means changing the identifier of the destination held at said header unit based on said configuration information.

[Claim 14] Said conversion means is switching equipment according to claim 13 characterized by changing said 1st logical address information into the 2nd logical address managed within the store subsystem used as the destination of this frame based on said configuration information held at said configuration management table including the 1st logical address information which shows the storing place of said data said frame is recognized to be by said data stereo section by said calculating machine.

[Claim 15] Said switching equipment is switching equipment according to claim 11 characterized by connecting with said transfer means, receiving the input of the configuration information which defines the configuration of the storage process defined system which consists of operators including this switching equipment and said two or more store subsystems, answering this input further, and having the management processor which sets said configuration information as the configuration management table of each node.

[Claim 16] Said 1st interface node is switching equipment according to claim 12 characterized by adding node address information which is different on each frame, and transmitting to said transfer means so that the light command frame which directs the writing of the data transmitted from said calculating machine may be answered, those duplicates may be generated about the data frame following this light command frame and it and the data frame following said light command frame and it may be sent to at least two store subsystems.

[Claim 17] Said 1st interface node is switching equipment according to claim 16 characterized by adding node address information which is different in each lead command frame, and transmitting to said transfer means so that the lead command frame which directs the lead of the data transmitted from said calculating machine may be answered, the duplicate of this lead command frame may be generated and said lead command frame may be sent to said at least two store subsystems.

[Claim 18] Said 1st interface node is switching equipment according to claim 17 characterized by receiving the data frame which answers said lead command frame and is transmitted from said at least two store subsystems, choosing one of these, and transmitting to said calculating machine.

[Claim 19] Said 1st interface node is switching equipment according to claim 16 characterized by answering the lead command frame which directs the lead of the data transmitted from said calculating machine, adding the node address information on the 2nd interface node linked to the store subsystem of 1 beforehand defined between said at least two store subsystems to said lead command frame, and transmitting to said transfer means.

[Claim 20] Two or more store subsystems which have the store which has a storage holding data, and the control device which controls this store, The 1st interface node connected to the computer which uses the data held at said two or more store subsystems, Two or more 2nd interface nodes by which each was connected to any one of said the store subsystems, A transfer means for said 1st interface node and said two or more 2nd interface nodes to be connected, and to transmit a frame between said 1st interface node and said two or more 2nd interface nodes, The store system characterized by having the management processor which connects with said transfer means, is equipped with the managed table holding the configuration information which defines the store structure of a system inputted by the operator, and manages this store structure of a system based on said configuration information.

[Translation done.]

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention] This invention relates to the approach of improvement in the speed of a disk control system, low-cost-izing, and improvement in cost performance especially about the implementation approach of the disk control system which controls two or more disk units.
[0002]
[Description of the Prior Art] There is a disk array system which controls two or more disk units as a store system used for a computer system. About a disk array system, it is "A Case for Redundant Arrays of Inexpensive Disks" (RAID), for example.; It is indicated by InProc.ACM SIGMOD and June 1988 (University of California at Berkeley issue). A disk array is operating two or more disk units to juxtaposition, and is the technique of realizing improvement in the speed compared with the storage system which used the disk unit alone.
[0003] There is an approach which used Fabric of a fiber channel (Fibre Channel) as an approach of connecting two or more disk array systems mutually with two or more hosts. the example of the computing system which applied this approach — Nikkei electronics 1995.7.3 (no.639) "Serial SCSI is to commercial scene still more" P.79 It is shown in drawing 3 . In the computing system indicated here, two or more host computers (below, it is only called a host) and two or more disk array systems are connected to fabric equipment through a fiber channel, respectively. Fabric equipment is the switch of a fiber channel and connects the transfer way between the equipment of the arbitration linked to fabric equipment. Fabric equipment is transparency to a transfer of the "frame" which is the packet of a fiber channel, and a host and a disk array system communicate in two points, without being conscious of fabric equipment of each other.
[0004]
[Problem(s) to be Solved by the Invention] In the conventional disk array system, the number of a disk unit is increased for large-capacity-izing, and if it is going to realize the controller which has the engine performance which balanced the number for high-performance-izing, the performance limit of the internal bus of a controller and the performance limit of the processor which performs transfer control will actualize. In order to cope with such a problem, an internal bus is extended and increasing the number of processors is performed. However, the method of such management causes the complication of control software and the increment in an overhead by complication of a controller configuration, exclusive control of the share data of interprocessor, etc. by much bus control. For this reason, while raising cost very much, the engine performance is reaching the ceiling, consequently cost performance gets worse. Moreover, in a large-scale system, although such equipment can realize the engine performance corresponding to the cost, there is a technical problem which does not balance that the increase of a development cycle and the rise of development cost to which expandability is restricted are caused in the system whose scale is not so large.
[0005] By putting two or more disk array systems in order, and interconnecting with fabric equipment, it is possible to perform large-capacity-izing as the whole system and high performance-ization. However, by this approach, since it cannot be distributed to other equipments even if it is irrelevant between disk array systems and access concentrates on a specific disk array system, high performance-ization on real use is unrealizable. Moreover, since the capacity of the logical disk unit (it is called Logical unit) seen from the host is restricted to the capacity of one set of a disk array system, large capacity-ization of Logical unit is unrealizable.
[0006] Although the mirror configuration by two sets of disk array systems can be realized using the mirroring function which the host has when it is going to form the whole disk array system into high reliance, the control overhead for mirroring by the host occurs, and the technical problem that system performance is restricted occurs. Moreover, if many disk array systems exist according to an individual in a

system, a load for a system administrator to manage will increase. For this reason, management cost increases — many maintenance staffs and the maintenance costs for two or more sets are needed. Furthermore, since two or more disk array systems and fabric equipment are isolated systems, respectively, it is necessary to carry out various setup by different approach for every equipment. For this reason, employment cost increases with a manager's training and increase of an operate time.

[0007] The purpose of this invention solves the technical problem in these conventional technique, can build the storage system according to the scale of a computing system, a demand, etc., and is to realize the storage system which can respond to the escape of the storage system in the future, improvement in dependability, etc. easily.

[0008]

[Means for Solving the Problem] The store which has the storage with which the store system of this invention holds data, Two or more store subsystems which have the control device which controls this store, The 1st interface node connected to the computer which uses the data held at two or more store subsystems, Two or more 2nd interface nodes by which each was connected to either of the store subsystems, And the 1st interface node and two or more 2nd interface nodes are connected, and it has a transfer means to transmit a frame between the 1st interface node and two or more 2nd interface nodes.

[0009] Preferably, the 1st interface node answers the frame sent from a calculating machine, analyzes this frame, carries out signal transduction to the configuration management table which stored store structure-of-a-system information about the destination of the frame based on the configuration information held at the configuration management table, and is transmitted to a transfer means.

[0010] Moreover, on the occasion of a transfer of a frame, the 1st interface node adds the node address information on the node which should receive the frame to a frame. A transfer means transmits a frame according to the node address information added to the frame. The 2nd interface node carries out the reconstititution of the frame except for node address information from the frame received from the transfer means, and transmits it to the target store subsystem.

[0011] In a mode with this invention, a storage system has a management processor linked to a transfer means. A management processor sets configuration information as a configuration management table according to the directions from an operator. The information which restricts access from a computer is included in configuration information.

[0012]

[Embodiment of the Invention] [1st operation gestalt] drawing 1 is a block diagram in 1 operation gestalt of the computer system using the disk array system by which this invention was applied.

[0013] It is the host computer (host) to which, as for 1, a disk array system is connected to, and, as for 30, a disk array system is connected. The disk array system 1 has the communication interface 80 between the disk array system-configuration-control means 70, and the disk array switch 20 and the disk array system-configuration-control means 70 of performing setting management of the disk array subset 10, the disk array switch 20, and the whole disk array system, and between the disk array subset 10 disk-array system-configuration-control means 70 (communication link I/F). It connects with the host interface (host I/F) 31, and a host 30 and the disk array system 1 connect host I/F31 to the disk array switch 20 of the disk array system 1. In the interior of the disk array system 1, the disk array switch 20 and the disk array subset 10 are connected with a disk array interface (disk array I/F21).

[0014] Although a host 30 and four disk array subsets 10 are shown respectively, they are [no limit] about this number and are arbitrary by a diagram. The number of a host 30 and the disk array subset 10 may differ. Moreover, the disk array switch 20 is doubled with this operation gestalt as illustration. Each host 30 and each disk array subset 10 are connected to the both sides of the disk array switch 20 doubled by respectively separate host I/F31 and disk array I/F21. This is for enabling access to the disk array system 1 from a host 30 by using another side, even if one disk array switch 20, host I/F31, or disk array I/F21 breaks down, and realizing high availability. However, such doubleness is not necessarily indispensable and is selectable according to the reliability level required of a system.

[0015] Drawing 2 is the block diagram showing the example of 1 configuration of the disk array subset 10. The high order adapter which 101 interprets the command from host system (host 10), carries out a cache hit mistake judging, and controls the data transfer between host system and a cache, the shared memory (it is called a cache and a shared memory below) in which 102 stores the cache for disk data-access improvement in the speed and the share data between multiprocessors, and 104 are two or more disk units stored in the disk array subset 10. 103 is a low order adapter which controls a disk unit 104 and controls the data transfer between a disk unit 104 and a cache. 106 is a disk array subset configuration management means, communicates through the disk array system-configuration-control means 70 and communication link I/F80 which manage the disk array system 1 whole, and manages a setup of a configuration parameter, the report of fault information, etc.

[0016] The high order adapter 101, a cache and a shared memory 102, and the low order adapter 103 are doubled,,respectively. Like doubleness of the above-mentioned disk array switch 20, this reason is for realizing the sex for Takayoshi, and is not indispensable. Moreover, either of the doubled low order adapters 103 of each disk unit 104 is also controllable. Although the same memory means is shared from a viewpoint of low-cost-izing to the cache and the shared memory with this operation gestalt, of course, these can also be dissociated.

[0017] The high order adapter 101 includes the high order bus 1012 which performs communication link between a cache and a shared memory 102, and a high order MPU 1010 and the disk array I/F controller 1011, and data transfer. [the high order MPU 1010 which performs control of the high order adapter 101, host system 1011, i.e., the disk array I/F controller which controls disk array I/F21 which is connection I/F with the disk array switch 20, and]

[0018] Although one disk array I/F controller 1011 is shown every high order adapter 101 by a diagram, two or more disk array I/F controllers 1011 may be formed to one high order adapter.

[0019] The low order adapter 103 includes the low order bus 1032 which performs communication link between a cache and a shared memory 102, and the low order MPU 1030 and the disk I/F controller 1031 which control disk I/F which is an interface with the low order MPU 1030 which performs control of the low order adapter 103, and a disk 104, and data transfer. [the disk I/F controller 1031, and]

[0020] Although four disk I/F controllers 1031 are shown every low order adapter 103 by a diagram, the number is arbitrary and can be changed according to the configuration and the number of a disk to connect of a disk array.

[0021] Drawing 3 is the block diagram showing the example of 1 configuration of the disk array switch 20. The management processor (MP) which is a processor to which 200 performs control and management of the whole disk array switch, the crossbar switch with which 201 constitutes the mutual switch path of nxn, the disk array I/F node in which 202 is prepared every disk array I/F21, the host I/F node in which 203 is prepared every host I/F31, and 204 are communication link controllers which perform the communication link between the disk array system-configuration-control means 70. I/F between clusters for the pass whose 2020 connects a crossbar switch 201 with the disk array I/F node 202, the pass on which 2030 connects a crossbar switch 201 with the host I/F node 203, and 2040 to connect with other disk array switches 20, and constitute a cluster, and 2050 are the pass for connecting a crossbar switch 201 with MP200.

[0022] Drawing 4 is the block diagram showing the structure of a crossbar switch 201. 2010 is a switching port (SWP) which are the pass 2020, 2030, and 2050 linked to a crossbar switch 201, and a port which connects I/F2040 between clusters. SWP2010 has the same structure altogether and performs switching control of the transfer path to other SWP(s) [SWP / a certain]. Although the transfer path is shown only about one SWP by a diagram, the same transfer path exists among all SWP(s).

[0023] Drawing 5 is the block diagram showing the example of 1 configuration of the host I/F node 203. With this operation gestalt, in order to explain concretely, it is assumed that it is what uses a fiber channel for both host I/F31 and disk array I/F21. Of course, it is also possible as host I/F31 and disk array I/F21 to apply interfaces other than a fiber channel. By using the same interface for both the host I/F node 203 and the disk array I/F node 202, both are made to the same structure. In this operation gestalt, it is constituted like the host I/F node 203 which also shows the disk array I/F node 202 in drawing. Below, the host I/F node 203 is explained to an example.

[0024] The retrieval processor which searches to which node 2021 transmits the received fiber channel frame (it is only called a frame below) (SP), 2022 is a host 30 (in the case of the disk array I/F node 202). The interface controller which transmits and receives a frame between the disk array subsets 10 (IC), The switching controller which changes based on the result with which SP2021 searched 2022 to the frame which IC2023 received (SC), The packet generation section packet-ized in the format that a crossbar switch 201 can be passed in order that 2024 may transmit the frame which SC2021 changed to other nodes (SPG), The frame buffer which stores temporarily the frame which 2025 received (FB), The exchange table which manages an exchange number for 2026 to identify the exchange (Exchange) which are two or more frame trains which corresponded to the disk array access request command (it is only called a command below) from one host (ET), 2027 is a disk array configuration management table (DCT) which stores the configuration information of two or more disk array subsets 10.

[0025] As for each configuration sections of all of the disk array switch 20, it is desirable on the engine performance to consist of hardware logic. However, if the engine performance called for can be satisfied, it is also possible to realize the function of SP2021 or SC2022 by the program control using a general-purpose processor.

[0026] Each disk array subset 10 has managed the disk unit 104 which each has as 1 or two or more logical disk units. This logical disk unit is called Logical unit (LU). LU does not need to correspond by the physical disk unit 104 and 1 to 1, and two or more LUs may be constituted by one set of a disk unit 104, or one LU may consist of two or more disk units 104.

[0027] When it sees from the outside of the disk array subset 10, one LU is recognized as one set of a disk unit. With this operation gestalt, still more logical LU is constituted by the disk array switch 20, and a host 30 operates so that it may access to this LU. On these specifications, when one LU recognized from a host 30 consists of an independent LU (ILU) and two or more LUs in LU recognized by the host 30 when one LU recognized from a host 30 consists of one LU, LU recognized by the host 30 is called Integration LU (CLU).

[0028] The correspondence relation of the address space between each hierarchy in case one integration LU is constituted from an LU of four disk array subsets by drawing 12 is shown. In drawing, the address space in one integration LU of the disk array system 1 which saw 1000 from the host "#2", and 1100 show the address space of LU of the disk array subset 10 as an example, and 1200 shows the address space of a disk unit 104 (here, illustrated only about the disk array subset "#0").

[0029] LU of each disk array subset 10 shall be constituted as a RAID5 (Redundant Arrays of Inexpensive Disks Level 5) mold disk array by four sets of disk units 104 here. Each disk array subset 10 has LU which has the capacity of n0, n1, n2, and n3, respectively. The disk array switch 20 unifies the address space which these four LUs have to the address space which has the capacity of (n0+n1+n2+n3), and integration LU recognized from a host 30 is realized.

[0030] With this operation gestalt, when host #2 access a field A1001, the access request which specified the field A1001 is changed into the demand for accessing the field A'1101 of LU of disk array subset #0 with the disk array switch 20, and is transmitted to disk array subset #0, for example. Disk array subset #0 accesses a field A'1101 further by mapping in 1201 A" of fields on a disk unit 104. Mapping between an address space 1000 and an address space 1100 is performed based on the configuration information held at DCT207 which the disk array switch 20 has. About the detail of this processing, it mentions later. In addition, it is the technique already well known about mapping in a disk array subset, and omits about detailed explanation on these specifications.

[0031] In this operation gestalt, DCT207 contains a system configuration table and a subset configuration table. Drawing 6 shows the configuration of a system configuration table, and drawing 7 shows the configuration of a subset configuration table.

[0032] As shown in drawing 7 , the system configuration table 20270 has the host LU configuration disk array I/F node configuration table 20271 and 20272 holding the information which shows Host's LU configuration showing the connection relation between the disk array I/F node 202 of the disk array switch 20, and the disk array subset 10.

[0033] The host LU configuration table 20271 has LU information (LU Info.) which is the information about LU of Host-LU No. which is the number which was seen from the host 30, and which identifies the LU for every LU, LU Type which shows the attribute of LU, CLU Class and CLU Stripe Size, Condition that is the information which shows Host's LU condition, and the disk array subset 10 which constitutes Host LU.

[0034] LU Type is information which shows the class of LU whether this host LU is CLU or to be ILU. CLU Class is information which shows any of "Joined", "mirrored", and "Striped" that class is, when it is shown by LU Type that this host LU is CLU. "Joined" shows that CLU which connects some LUs and has one big storage space is constituted, as drawing 11 explained. "Mirrored" shows that it is LU doubled by two LUs so that it may mention later as the 6th operation gestalt. "Striped" consists of two or more LUs, and shows that it is LU by which data were distributed and stored in LU of these plurality so that it may mention later as the 7th operation gestalt. CLU Stripe Size shows striping size (size of the block used as the unit of distribution of data), when it is shown by CLU Class that it is "Striped."

[0035] There are four kinds of the conditions by which it is shown by Condition, "Normal", "Warning", "Fault", and "Not Defined." As for "Normal", this host LU shows that it is in a normal condition. "Warning" shows that degeneration operation is performed to one corresponding to LU which constitutes this host LU of disk units for the reason of the failure having occurred. "Fault" shows that this host LU cannot be operated by failure of the disk array subset 10 etc. "Not Defined" shows that the corresponding host LU of Host-LU No. is not defined.

[0036] LU Info includes LUN within the information which specifies the disk array subset 10 to which that LU belongs about LU which constitutes this host LU, and a disk array subset, and the information which shows that size. When Host LU is ILU, the information about the only LU is registered. When Host LU is CLU, the information about each LU is registered about all LUs that constitute it. For example, in drawing, it is CLU which consists of four LUs, LUN "0" of a disk array subset "#0", LUN "0" of a disk array subset "#1", LUN "0" of a disk array subset "#2", and LUN "0" of a disk array subset "#3", and, as for Host-LU whose Host-LU No. is "0", it turns out that it is CLU the CLU class of whose is "Joined."

[0037] The disk array I/F node configuration table 20272 holds the information which shows of which disk array switch 20 the disk array I/F node 202 is connected for every port of the disk array subset 10 which 'disk array I/F21 connects.

[0038] Specifically, it has Subset No. which specifies the disk array subset 10, Subset Port No. which pinpoints a port, Switch No. which specifies the disk array switch 20 linked to the port, and I/F Node No. which specifies the disk array I/F node 202 of the disk array switch 20. When the disk array subset 10 is equipped with two or more ports, information is set up for every port of the.

[0039] A subset configuration table has two or more tables 202720-202723 corresponding to each disk array subset 10, as shown in drawing 7 . Each table contains LU configuration table 202740 holding the information which indicates the configuration of LU built in the disk array subset 10 to be the RAID group configuration table 202730 holding the information which shows a RAID group's configuration built within the disk array subset 10.

[0040] In a configuration of that striping of the RAID level 0 and the 5 grades was carried out, Group No. which shows the number by which the RAID group configuration table 202730 was added to the RAID group, Level which shows the RAID group's level, Disks which is the information which shows the number of the disks which constitute the RAID group, and its RAID group contain as information Stripe Size which shows the stripe size. For example, in the table shown in drawing, a RAID group "0" is a RAID group constituted by four sets of disk units, RAID level is 5 and stripe size is S0.

[0041] LU configuration table 202740 contains as information LU No. which shows the number (LUN) added to LU, RAID Group which shows whether this LU is constituted by which RAID group, Condition which shows the condition of LU, Size which shows the size (capacity) of this LU, Port which shows whether this LU is accessible from the port of disk array subset 10 throat, and Alt.Port which shows the port used as that alternative. The condition by which it is shown by Condition has four kinds, "Normal", "Warning", "Fault", and "Not Defined", like Condition about Host LU. When a failure occurs in the port pinpointed for the information set as Port, it is used, but the port pinpointed using the information set as Alt.Port can also be used in order to only access the same LU from two or more ports.

[0042] Drawing 8 is the block diagram of the frame in a fiber channel. The frame 40 of a fiber channel contains CRC (Cyclic RedundancyCheck)403 which is an error detection code with a frame payload of 402 or 32 bits which is the part which stores SOF (Start Of Frame)400, the frame header 401, and the actual condition data of a transfer in which the head of a frame is shown, and EOF (End Of Frame)404 which shows the tail end of a frame. The frame header 401 has structure as shown in drawing 9 , and ID (SEQ_ID) of a sequence which specifies the frame group under each exchange ID specified by a response place (OX_ID, RX_ID) and exchange is stored the ID [of the frame source] (S_ID), ID [of the frame destination] (D_ID), and starting origin of an exchange.

[0043] With this operation gestalt, ID by which ID assigned to the host 30 as S_ID was assigned as D_ID to the port of the disk array switch 20 again is used for the frame published by the host 30. The exchange ID of one pair (OX_ID, RX_ID) is assigned to one host command. When it is necessary to publish two or more data frames to the same exchange, the same SEQ_ID is assigned to all the data frames, and each is identified at a sequence count (SEQ_CNT). The maximum length of the frame payload 402 is 2110 bytes, and the contents stored for every frame kind differ. For example, in the case of the FCP_CMD frame mentioned later, as shown in drawing 10 , SCSI Logical Unit Number (LUN), Command Description Block (CDB), etc. are stored. CDB contains a command cutting tool required for disk (disk array) access, the transfer initiation logical address (LBA), and transfer length (LEN).

[0044] Hereafter, actuation of the disk array system of this operation gestalt is explained.

[0045] Before using a disk array system, it is necessary to the disk array switch 20 to set up the configuration information of the disk array subset 10. A system administrator acquires all the disk array subsets 10 and the configuration setting information on the disk array switch 20 from an administration terminal 5 through the disk array system configuration means 70. A manager inputs setting information required for various setup, such as a configuration setup of Logical unit, a setup of RAID level, and a setup of the shift pass at the time of failure generating, so that it may become a desired system configuration from an administration terminal 5. The disk array system-configuration-control means 70 receives the setting information, and transmits setting information to each disk array subset 10 and the disk array switch 20. In addition, the 5th operation gestalt explains separately the input of the setting information in an administration terminal 5.

[0046] With the disk array switch 20, the communication link controller 204 acquires setting information, and configuration information, such as address space information on each disk array subset 10, is set up by MP200. MP200 distributes the configuration information of the disk array subset 10 to each host I/F node 203 and the disk array I/F node 202 by crossbar switch 201 course.

[0047] Each nodes 203 and 202 store configuration information in DCT2027 by SP2021, if this information is received. In the disk array subset 10, the disk array subset configuration management means 106 acquires setting information, and stores in a shared memory 102. Each high order MPU 1010 and low order MPU 1030 carry out each configuration management with reference to the setting information on a shared memory 102.

[0048] Below, actuation when a host "#2" publishes a lead command to the disk array system 1 is explained. The flow chart of actuation [in / for the mimetic diagram showing the sequence of the frame transmitted to drawing 11 through a fiber channel at the time of the lead actuation from a host / the host I/F node 203 of the disk array switch at this time] is shown in drawing 13 .

[0049] In addition, in the following explanation, a host "#2" assumes accessing the storage region A1001 in drawing 12 . A " of actual storage regions corresponding to a storage region A1001 shall exist in the address space of disk unit #2 which constitute LU of LUN=0 of a disk array subset "#0." Moreover, "CLU" shall be set to LU Type of the host LU configuration table 20271 which defines LU which constitutes an address space 1000, and "Joined" shall be set to CLU Class.

[0050] A host 30 publishes the command frame "FCP_CMD" which stored the lead command on the disk array switch 20 at the time of the lead of data (drawing 11 arrow head (a)). The host I/F node "#2" of the disk array switch 20 receives a command frame "FCP_CMD" by host I/F31 course by IC2023 (step 20001). IC2023 transmits a command frame to SC2022. SC2022 once stores the received command frame in FB2025. Under the present circumstances, SC2022 calculates CRC of a command frame and inspects that receipt information is right. If an error is in inspection of CRC, SC2022 will notify that to IC2023. IC2023 will report a CRC error to a host 30 through host I/F31, if the notice of an error is received from SC2022. (Step 20002) .

[0051] The frame to which CRC held SC2022 to FB2025 the right case is led, it recognizes that it is a command frame, and the frame header 401 is analyzed (step 20003). And SC2022 is directed to SP2021 and registers exchange information, such as S_ID, D_ID, and OX_ID, into ET2026 (step 20004).

[0052] Next, SC2022 analyzes the frame payload 402 and acquires LUN and CDB which were specified by the host 30 (step 20005). With directions of SC2022, SP2021 searches DCT2027 and gets the configuration information of the disk array subset 10. Specifically, SP2021 finds the information which has Host-LU No. which is in agreement with LUN stored in the frame payload 402 which searched the host LU configuration table 20271 and was received. SP2021 recognizes Host's LU configuration from the information set as LU Type and CLU Class, and distinguishes LUN of the disk subset 10 which should be accessed based on the information currently held at LU Info., and LU in it, and LBA within this LU. Next, with reference to LU configuration table 202740 of the subset configuration table 202720, SP2021 checks the connection port of the target disk array subset 10, and gets node No. of the disk array I/F node 202 linked to the port from the disk array I/F node configuration table 20272. SP2021 reports the conversion information that the disk array subset 10 which carried out in this way and was obtained is identified, such as a number, and LUN, LBA, to SC2022. (Step 20006) .

[0053] Next, SC2022 changes LBA in LUN and CDB of the frame payload 402 using the acquired conversion information. Moreover, D_ID of the frame header 401 is changed into D_ID of the corresponding host I/F controller 1011 of the disk array subset 10. In addition, S_ID is not rewritten at this time (step 20007).

[0054] SC2022 transmits the disk array I/F node number linked to the command frame and the object disk array subset 10 after conversion to SPG2024. SPG2024 generates the packet which added the easy extended header 601 as shown in drawing 14 to the command frame after the received conversion. This packet is called the switching packet (S Packet) 60. In the extended header 601 of SPacket60, a source (self-node) number, a destination node number, and transfer length are ****** rare **. SPG2024 transmits generated S Packet60 to a crossbar switch 201 (step 20008).

[0055] A crossbar switch 201 receives S Packet60 by SWP2010 linked to a host I/F node "#2." SWP2010 performs switch control to SWP which the node of the destination connects with reference to the extended header 601 of S Packet60, establishes a path, and transmits S Packet60 to the disk array I/F node 202 (here disk array I/F node "#0") of the destination. SWP2010 will release the path, if it carries out at every reception of establishment of a path of S Packet60 and a transfer of S Packet60 is completed. In a disk array I/F node "#0", SPG2024 receives S Packet60, removes the extended header 601, and passes the part of a command frame to SC2022.

[0056] SC2022 writes its ID in S_ID of the frame header of the received command frame. Next, to SP2021, it points to SC2022 so that exchange information, such as S_ID of a command frame, D_ID, and OX_ID, and a frame source host I/F node number may be registered into ET2026, and it transmits a command frame to IC2023. IC2023 transmits a command frame to the disk array subset 10 (here disk array subset "#0") to connect according to the information on the frame header 401 (drawing 11 arrow head (b)).

[0057] A disk array subset "#0" receives the command frame after conversion "FCP_CMD" by the disk array I/F controller 1011. It is recognized as a high order MPU 1010 being a command which acquires LUN and CDB which were stored in the frame payload 402 of a command frame, and leads the data of LEN length from LBA of the specified Logical unit.

[0058] A high order MPU 1010 performs cache hit mistake / hit judging with reference to the cache management information stored in the shared memory 102. If it hits, data transfer will be carried out from a cache 102. Since it is necessary to lead data from a disk unit in a mistake, address translation based on the configuration of RAID5 is carried out, and cache space is secured. And processing information required for lead processing from a disk unit 2 is generated, and processing information is stored in a shared memory 102 in order to take over processing to low order MPU 1030.

[0059] Low order MPU 1030 starts processing ignited by processing information having been stored in the shared memory 102. Low order MPU 1030 specifies the suitable disk I/F controller 1031, generates the lead command to a disk unit 2, and publishes a command for the disk I/F controller 1031. The disk I/F controller 1031 stores the data led from the disk unit 2 in the address with which the cache 102 was specified, and notifies a termination report to low order MPU 1030. Low order MPU 1030 stores processing termination information in a shared memory 102 that it should notify that processing was completed correctly to a high order MPU 1010.

[0060] A high order MPU 1010 resumes processing ignited by processing termination information having been stored in the shared memory 102, and notifies lead data-preparation completion to the disk array I/F controller 1011. The disk array I/F controller 1011 publishes "FCP_XFER_RDY" which is a data transfer preparation-completion frame in a fiber channel to the disk array I/F node "#0" concerned of the disk array switch 20 (drawing 11 arrow head (c)).

[0061] In a disk array I/F node "#0", if a data transfer preparation-completion frame "FCP_XFER_RDY" is received, SC2022 will gain the response place exchange ID (RX_ID) which received from the disk array subset 20, will specify S_ID, D_ID, and OX_ID, will direct to SP2021, and will register RX_ID into the exchange information concerned on ET2026. SC2022 gains the host I/F node number of the destination (source of a command frame) of a data transfer preparation-completion frame. SC2022 cancels S_ID of this frame and transmits it to SPG2024. As SPG2024 was described previously, it generates S Packet, and it transmits it to a candidate host I/F node "#2" by crossbar switch 201 course.

[0062] In a host I/F node "#2", if SPG2024 receives S Packet of a data transfer preparation-completion frame, the extended header of S Packet will be removed, "FCP_XFER_RDY" will be reproduced, and SC2022 will be passed (step 20011). SC2022 is directed to SP2021 and specifies the exchange which searchs ET2026 and corresponds (step 20012).

[0063] Next, SC2022 investigates whether a frame is "FCP_XFER_RDY" (step 20013), and if it is "FCP_XFER_EDY", it will direct renewal of the response place exchange ID of ET2026 (RX_ID) to SP2021. The value added to this frame is used as response place exchange ID (step 20014). And SC2022 changes S_ID of the frame header 401, and D_ID into the suitable value which used ID of the host I/F node 203, and a host's 30 ID (step 20015). The frame header 401 is changed into the frame to a host "#2" by these processings. IC2023 publishes this data transfer preparation-completion frame "FCP_XFER_RDY" to a host "#2" (the arrow head of drawing 11 (d): step 20016).

[0064] In order that the disk array I/F controller 1011 of a disk array subset "#0" may perform data transfer, it generates a data frame "FCP_DATA" and transmits it to the disk array switch 20 (drawing 11 arrow head (e)). Since the transfer length of a frame payload has a limit, the greatest data length which can be transmitted by one frame is 2KB. When a data length exceeds this, only a required number generates and publishes a data frame. The same SEQ_ID is assigned to all data frames. Issue of a data frame is the same as that of the case where it is a data transfer preparation-completion frame except for two or more frames being generated to the same SEQ_ID (that is, SEQ_CNT changing).

[0065] The disk array switch 20 changes the frame header 401 of a data frame "FCP_DATA" like processing of a data transfer preparation-completion frame. However, since RX_ID is already established in the transfer of a data frame, processing of step 20014 in processing of a data transfer preparation-completion frame is skipped. The disk array switch 20 transmits a data frame to a host "#2" after conversion of the frame header 401 (drawing 11 arrow head (f)).

[0066] Next, in order that the disk array I/F controller 1011 of a disk array subset "#0" may perform an exit-status transfer, it generates a status frame "FCP_RSP" and publishes it to the disk array switch 20 (drawing 11 arrow head (g)). With the disk array switch 20, like processing of a data transfer preparation-completion frame, SPG2024 removes an extended header from S Packet, reproduces a "FCP_RSP" status frame (step 20021), searches ET2026 by SP2021, and acquires exchange information (step 20022). SC2022 changes a frame based on the information (20023 of a step). The changed frame is transmitted to a host "#2" by IC2023 (drawing 11 arrow head (h): step 20024). Finally SP2021 deletes exchange information from ET2026 (step 20025).

[0067] Lead processing from a disk array is performed as mentioned above. Lead processing mentioned above only by the direction of transfer of a data frame being reversed also about the light processing to the disk array system 1 and same processing are performed.

[0068] As shown in drawing 3 , the disk array switch 20 equips the crossbar switch 201 with I/F2040 between clusters. I/F2040 between clusters is not used in the system configuration shown in drawing 1 . The disk array switch 20 of this operation gestalt is mutually connectable with other disk array switches, as shown in drawing 15 using I/F2040 between clusters.

[0069] the disk array switch 20 in this operation gestalt — although the disk array subset 10 is connectable only to a total of eight sets with a host 30 if independent, the number of disk arrays can be increased with the host 10 who interconnects and can connect two or more disk array switches using I/F2040 between clusters. For example, in the system shown in drawing 15 , the disk array subset 10 can be connected to a total of 32 sets with a host 30 using four sets of the disk array switches 20, and data transfer becomes possible mutually among these.

[0070] Thus, with this operation gestalt, the connection number of a disk array subset or a host can be increased according to the need for disk capacity or the engine performance. Moreover, since between host-disk array systems is connectable using host I/F for a required transfer band, the expandability of capacity, the engine performance, and the number of connection can be raised sharply.

[0071] According to the operation gestalt explained above, even if the engine performance of one set of a disk array subset is restricted by internal MPU and an internal internal bus, it can interconnect between disk array subsets with a host with a disk array switch using two or more disk array subsets. Thereby, the engine performance high as a disk array system total is realizable. Even if the engine performance of a disk array subset is comparatively low, high performance-ization is realizable by using two or more disk array subsets. Therefore, according to the scale of a computer system, only the required number can connect the disk array subset of low cost, and it becomes possible to build a disk array system at the suitable cost according to a scale.

[0072] Moreover, what is necessary is just to add a disk array subset as required when increase of disk capacity and improvement in the engine performance are needed. Furthermore, since the host and disk array subset of a number of arbitration are connectable using two or more disk array switches, both capacity the engine performance and the number of connection can be raised sharply, and the system which has high expandability can be realized.

[0073] Since the contraction machine of the conventional disk array system itself can be used as a disk array subset according to this operation gestalt, the already developed large-scale control-software property can be used as it is, and reduction of development cost and compaction of a development cycle can be realized further again.

[0074] [2nd operation gestalt] drawing 16 is the block diagram of the computer system in the 2nd operation gestalt of this invention. This operation gestalt changes only the frame header 401 in the host I/F node of a disk array switch, and the frame payload 402 is different from the 1st operation gestalt constitutionally at the point which the point which is not operated and a disk array switch, host I/F, and disk array I/F have not doubled. Therefore, the configuration of each part does not have the 1st operation gestalt and the place which changes a lot, and omits explanation about the detail.

[0075] In drawing 16 , each disk array subset 10 consists of two or more Logical unit (LU) 110. Each LU110 is constituted as an independent LU. Generally, LUN assigned to LU110 in each disk array subset 10 is the consecutive number which begins from 0. For this reason, to show continuously LUN of all LUs110 in the disk array system 1 to a host 30, it is necessary to change the LUN field of the frame payload 402 like the 1st operation gestalt. With this operation gestalt, by showing a host 30 LUN of each disk array subset 10 as it is, conversion of the frame payload 402 is made unnecessary and control of a disk array switch is made easy.

[0076] The disk array switch 20 of this operation gestalt is assumed to be what can access the specific disk array subset 10 every host I/F node 203. In this case, if one host I/F31 is used, only LU110 in one set of the disk array subset 10 is accessible. The host is connected to two or more host I/F nodes 203 to access LU110 of two or more disk array subsets 10 from one set of a host. Moreover, when enabling it to access LU110 of one set of the disk array subset 10 from two or more hosts 30, loop-formation topology, fabric topology, etc. are used for the same host I/F node 203, and two or more hosts 30 are connected. Thus, since the disk array subset 10 will be decided for every D_ID of the host I/F node 203 in case one LU110 is accessed from one set of a host 30 if constituted, it is possible to show a host 30 LUN of each LU as it is.

[0077] With this operation gestalt, for the reason mentioned above, since LUN of LU110 in each disk array subset 10 is shown to the host 30 as it is at the host 30, conversion of LUN in the disk array switch 20 becomes unnecessary. For this reason, the disk array switch 20 will change only the frame header 401 like the 1st example, if a frame is received from a host 30, and it transmits the frame payload 402 to the disk array subset 10, without changing. Since actuation of each part in this operation gestalt is the same as that of the 1st operation gestalt if it removes that conversion of the frame payload 402 is not performed, detailed explanation is omitted here. According to this operation gestalt, development of the disk array switch 20 can be made easy.

[0078] With the 2nd operation gestalt of the [3rd operation gestalt], in the host I/F node of a disk array switch, although only the frame header is changed, by the 3rd operation gestalt explained below, the gestalt which does not change a frame is explained also including a frame header. The computer system of this operation gestalt is constituted like the computer system in the 1st operation gestalt shown in drawing 1 .

[0079] With the 1st and 2nd operation gestalt, internal configurations of the disk array system 1, such as the number of the disk array subset 10 and a configuration of LU110, are concealed to a host 30. For this reason, from a host 30, the disk array system 1 appears as one storage on the whole. On the other hand, the disk array subset 10 is opened to a host 30 as it is, and a host 30 enables it to use ID of the port of a direct disk array subset as D_ID of a frame header with this operation gestalt. Thereby, a disk array switch just needs to control a transfer of a frame according to the information on a frame header, and can change and use switching equipment equivalent to the fabric equipment of the fiber channel in the conventional technique for the disk array switch 20.

[0080] The disk array system-configuration-control means 70 communicates with the means of communications 204 of the communication link controller 106 of the disk array subset 10, and the disk array switch 20, and gains or sets up the configuration information of each disk array subset 10 and the disk array switch 20.

[0081] The disk array switch 20 has fundamentally the same configuration as the disk array switch in the 1st operation gestalt shown in drawing 3 . However, in order to control a transfer of a frame by this operation gestalt, using the information on the frame header of the frame which a host 30 publishes as it is, the function of conversion, such as DCT2027 which the host I/F node 203 of the disk array switch 20 and the disk array I/F node 202 have with the 1st operation gestalt or the 2nd operation gestalt, and a frame header realized by SC2022 and SPG2024 grade, becomes unnecessary. The crossbar switch 201 which the disk array switch 20 has transmits the frame of a fiber channel according to the information on a frame header between the host I/F node 203 and the disk array I/F node 202.

[0082] With this operation gestalt, in order to manage the disk array structure of a system collectively with the disk array system-configuration-control means 70, the disk array system-configuration-control means 70 is equipped with a disk array administrative table (this table is also hereafter called DCT). DCT with which the disk array system-configuration-control means 70 is equipped contains two table groups, the system configuration table 20270 shown in drawing 6 and 7, and the subset configuration tables 202720-202723. In addition, with this operation gestalt, since Host LU is altogether constituted as ILU, all LU Type (s) of the host LU configuration table 20271 serve as "ILU", and CLU Class and CLU Stripe Size do not make semantics.

[0083] A manager operates an administration terminal 5, communicates with the disk array system-configuration-control means 70, acquires information, such as disk capacity of the disk array subset 10, and the number of a disk unit, and performs a setup of LU110 of the disk array subset 10, a setup of RAID level, etc. Next, a manager communicates with the disk array system-configuration-control means 70 with an administration terminal 5, controls the disk array switch 20, and sets up the related information between the disk array subsets 20 with each host 30.

[0084] LU110 can come to be seen as the configuration of the disk array system 1 is established and a manager wishes from a host 30 by the above actuation. The disk array configuration management means 70 can save the above setting information, and according to the actuation from a manager, the check of a configuration is performed and it can make a change of a configuration.

[0085] Once it constitutes the disk array system 1 according to this operation gestalt, existence of the disk array switch 20 cannot be made to be able to recognize from a manager, and two or more disk array subsystems can be treated like one set of a disk array system. Moreover, according to this operation gestalt, the disk array switch 20 and the disk array subset 10 can be systematically operated according to the same operating environment, and the configuration check and a configuration change also become easy. Furthermore, without changing a setup of a host 30, when transposing the disk array system which was being used conventionally to the disk array system in this operation gestalt according to this operation gestalt, it can double with the disk array structure of a system which was using the configuration of the disk array system 1 till then, and compatibility can be maintained.

[0086] With the 1st to 3rd operation gestalt explained beyond the [4th operation gestalt], the fiber channel is used for host I/F. The operation gestalt explained below explains the gestalt in which interfaces other than a fiber channel were intermingled.

[0087] Drawing 17 shows the example of 1 configuration of IC2023 of the host I/F node 203 interior in case host I/F is Parallel SCSI. The SCSI protocol controller (SPC) by which 20230 performs protocol control.qf Parallel SCSI, the fiber channel protocol controller (FPC) by which 20233 performs protocol control of a fiber channel, the protocol conversion processor (PEP) to which 20231 carries out protocol conversion of Serial SCSI of a fiber channel to Parallel SCSI, and 20232 are buffers (BUF) which save the data in protocol conversion temporarily.

[0088] In this operation gestalt, a host 30 publishes the SCSI command to the disk array I/F node 203. In the case of a lead command, SPC20230 stores this in BUF20232, and reports reception of a command to PEP20231 by interruption. PEP20231 uses the command stored in BUF20232, changes it into the command to FPC20233, and is sent to FPC20233. If this command is received, FPC20233 will be changed into a frame format and will be handed over to SC2022. Under the present circumstances, Exchange ID, Sequence ID, Source ID, and Destination ID are added by PEP20231 so that subsequent processings may be possible. Next command processing is performed like the 1st operation gestalt.

[0089] The disk array subset 10 will carry out issue of issue of a data transfer preparation-completion frame, data transfer, and the status frame after normal termination, if preparation of data is completed. While the frame header 401 and the frame payload 402 are changed if needed, a transfer of various frames is performed from the disk array subset 10 before IC2023. FPC20233 of IC2023 receives a data transfer preparation-completion frame, it receives data continuously, stores them in BUF20232, and if a transfer finishes normally continuously, it will report data transfer completion, it receiving a status frame and applying interruption to PTP20231. If interruption is received, PTP20231 starts SPC20230 and directs that it starts data transfer to a host 30. If SPC20230 transmits data to a host 30 and normal termination is checked, it will report normal termination by interruption to PTP20231.

[0090] Here, although Parallel SCSI was shown as an example of host I/F other than a fiber channel, it is possible to apply similarly to ESCON which are other interfaces, for example, host I/F to a mainframe. It is possible to make both so-called open systems, such as a mainframe, and a personal computer, a workstation, intermingled in one set of the disk array system 1, and to connect with it by preparing for example, a fiber channel and the host I/F node corresponding to Parallel SCSI and ESCON as a host I/F node 203 of the disk array switch 20. Although the fiber channel is used like the 1st to 3rd operation gestalt as disk array I/F with this operation gestalt, it is possible to use I/F of arbitration also to disk array I/F.

[0091] The approach of the configuration management of the [5th operation gestalt], next the disk array system 1 is explained as the 5th operation gestalt. Drawing 18 is the system configuration Fig. of this operation gestalt. With this operation gestalt, the host 30 is formed four sets. I/F30 between the disk array systems 1 is connected with a host "#0" and "#1" by Parallel SCSI (Ultra2 SCSI) between the fiber channel, the host "#2", and the disk array system 1 between Parallel SCSI (Ultra SCSI), and a host "#3" and the disk array system 1.

[0092] Connection with the disk array switch 20 of Parallel SCSI is made like the 4th operation gestalt. The disk array system 1 has four sets of the disk array subsets 30. Two independent LUs are constituted by four independent LUs and the disk array subset "#1" at the disk array subset "#0", respectively. One integration LU consists of a disk array subset "#2" and "#3." With this operation gestalt, like the 1st operation gestalt, the disk array subset 10 shall be concealed to a host 30, and the frame of a fiber channel shall be changed. LUN assigned to each LU — the order from LU of a disk array subset "#0" — LUN=0, 1 and 2, and ... they are 7 **s to 6.

[0093] Drawing 19 is an example of the screen displayed on the display screen of an administration terminal 5. Drawing is the logical connection configuration screen in which correspondence with host I/F31 and each Logical unit (LU) was shown.

[0094] The relation between the information 3100 about each host I/F31, the information 11000 about each LU110, the disk array subset 10, and LU110 etc. is displayed on the logical connection configuration screen 50. An I/F class, an I/F rate, the status, etc. are contained as information about host I/F31. As information about LU110, a storing subset number, LUN, capacity, RAID level, the status, information, etc. are displayed. By referring to this screen, a manager can manage the configuration of the disk array system 1 easily.

[0095] On the logical connection configuration screen 50, the line currently drawn between host I/F and LU shows accessible LU110 via each host I/F31. It cannot access from the host 30 who connects with the host I/F from host I/F to LU110 by which a line is not drawn. Since the data format to treat differs and a user also changes with hosts 30, it is indispensable on security maintenance to prepare a suitable access restriction. Then, the manager who sets up a system carries out access restriction by whether the access permission between each LU110 and host I/F is given using this screen. In drawing, although LU "#0" is accessible from host I/F "#0" and "#1", it cannot be accessed from host I/F "#2" and "#3." LU "#4" is accessible only from host I/F "#2."

[0096] In order to realize such an access restriction, access-restriction information is transmitted from the disk array system-configuration-control means 70 to the disk array switch 20. The access-restriction information sent to the disk array switch 20 is distributed to each host I/F node 203, and is registered into DCT2027 of each host I/F node 203. When the inspection command of LU existence existence to LU to which access was restricted is published by the host, it is each host I/F node's 203 inspecting DCT2027, and not answering to an inspection command, or returning an error, and the LU is no longer recognized from a host by him. In the case of a SCSI protocol, generally as an inspection command of LU existence existence, the Test Unit Ready command and the Inquiry command are used. Since read/write is not carried out without this inspection, it is possible to apply a limit of access easily.

[0097] Although access restriction is applied every host I/F31 with this operation gestalt, it is also easily realizable by extending this to apply access restriction every host 30. Moreover, host I/F31, a host 30, or an address space can be pinpointed, and the access restriction according to the classification of a command to which only the lead was told as good and which good, read/write, and good and read/write told that only a light was improper can also be applied. In this case, a host I/F number, Host ID, an address space, a limit command, etc. are specified as access-restriction information, and a limit is set as the disk array switch 20.

[0098] Next, the addition of the new disk array subset 10 is explained. When adding the disk array subset 10 newly, a manager connects the disk array subset 10 added to the disk array I/F node 202 as for which the disk array switch 20 is vacant. Continuously, a manager operates an administration terminal 5 and does the depression of the carbon button 5001 "reflecting the newest condition" currently displayed on the logical connection configuration screen 50. This actuation is answered and the picture showing a non-set up disk array subset is displayed on a screen (not shown). If it carries out by choosing the picture of this disk array subset, the setting screen of a disk array subset will appear. A manager carries out various setup of the disk array subset added newly on the displayed setting screen. There are a configuration of LU, RAID level, etc. in the item set up here. Continuously, if it changes to the screen of the logical connection block diagram of drawing 19 , a new disk array subset and LU will appear. Henceforth, if the access restriction which receives every host I/F31 is set up and the depression of the "setting activation" carbon button 5002 is carried out, to the disk array switch 20, the information on access-restriction information and a disk array subset, and LU will be transmitted, and a setup will be performed.

[0099] It is performed by the procedure which also mentioned above the procedure at the time of adding LU110 to each disk array subset 10. Moreover, it is performed by a disk array subset and the procedure with the same almost said of deletion of LU. A different point is a point performed, after a manager chooses each deletion part on a screen, and pushes "deletion" carbon button 5003 and a suitable check is performed. As mentioned above, a manager can manage the whole disk array system unitary by using an administration terminal 70.

[0100] Processing of the [6th operation gestalt], next mirroring by the disk array switch 20 is explained as the 6th operation gestalt. Mirroring explained here is the approach of supporting duplex writing by two independent LUs of two sets of disk array subsets, and is the doubleness which even the controller of a disk array subset included. Therefore, dependability differs from doubleness of only a disk.

[0101] The structure of a system in this operation gestalt is the same as what is shown in drawing 1 . the configuration shown in drawing 1 — it shall be, a disk array subset "#0" and "#1" shall completely be equipped with the same LU configuration, and these two disk array subsets shall be seen as one disk array from a host 30 For convenience, the number of the pair of the disk array subset by which mirroring was carried out is called "#01." Moreover, a mirroring pair is formed of LU "#0" and LU "#1" of each disk array subset, and the pair of this LU is called LU "#01" for convenience by them. "Mirrored" is set as CLU Class and, as for the information for managing LU#01 on the host LU configuration table 20271 of DCT2027, the information about LU#0 and LU#1 is set up as LU Info. The configuration of other each part is the same as that of the 1st operation gestalt.

[0102] Actuation of each part in this operation gestalt is the same as that of the 1st example almost. Hereafter, the point which is different from the 1st operation gestalt is explained focusing on actuation of the host I/F node 203 of the disk array switch 20. The mimetic diagram showing the sequence of the frame to which drawing 20 is transmitted at the time of the light actuation in this operation gestalt, drawing 21 , and 22 are flow charts which show the flow of processing by the host I/F node 203 at the time of light actuation.

[0103] The light command frame (FCP_CMD) which the host 30 published is received by IC2023 at the time of light actuation (the arrow head of drawing 20 (a): step 21001). The light command frame received by IC2023 is step 20002 at the time of the lead actuation explained with the 1st operation gestalt. It is processed like 20005 (step 21002-21005).

[0104] SC2022 searches DCT2027 using SP2021, and recognizes that it is a light access request to LU

"#01" of the mirror-ized disk array subset "#01" (step 21006). SC2022 creates the duplicate of the command frame which received on FB2025 (step 21007). SC2022 changes a command frame based on the configuration information set as DCT2027, and creates the separate command frame of both LU "#0" and LU "#1" (step 21008). Here, main LU and LU"#1" is called a main command frame and ** command frame for LU "#0" also to ** LU, a call, and a command frame, respectively. and both -- exchange information is separately stored in ET2026, and the command frame created to the disk array subset "#0" and the disk array subset "#1" is published (the arrow head of drawing 20 (b0) (b1): step 21009).

[0105] Each disk array subset "#0" and "#1" receive a command frame, and they transmit a data transfer preparation-completion frame (FCP_XFER_RDY) to the disk array switch 20 independently, respectively (arrow head of drawing 20 (c0) (c1)). Step 20011 of lead actuation [in / by the disk array switch 20 / in the host I/F node 203 / the 1st operation gestalt] The data transfer preparation-completion frame transmitted by the same processing as 20013 is processed (step 21011-21013).

[0106] In the phase which had complete set of data transfer preparation-completion frame from each disk array subset, (step 21014) and SC2022 carry out conversion to the main data transfer preparation-completion frame (step 21015), and transmit the frame after conversion to a host 30 by IC2023 (the arrow head of drawing 20 (d): step 21015).

[0107] A host 30 transmits a data frame (FCP_DATA) to the disk array switch 20 for light data transmission, after receiving a data transfer preparation-completion frame (arrow head of drawing 20 (e)). If the data frame from a host 30 is received by IC2023 (step 21031), like a lead command frame or a light command frame, it will be stored in FB2025 and CRC inspection and analysis of a frame header will be performed (steps 21032 and 21033). Based on the analysis result of a frame header, ET2026 is searched by SP2021 and exchange information is acquired (step 21034).

[0108] SC2022 creates a duplicate like the time of a light command frame (step 21035), one of these is turned to LU "#0" in a disk array subset "#0", and turns another side to LU "#1" in a disk array subset "#1", and transmits (the arrow head of drawing 20 (f0) (f1): step 21037).

[0109] A disk array subset "#0" and "#1" receive a data frame, it carries out a light to a disk unit 104, respectively, and they transmit a status frame (FCP_RSP) to the disk array switch 20.

[0110] SC2022 -- a disk array subset "#0" and "#1" -- respectively -- since -- if a status frame is received, an extended header will be removed from those status frames, a frame header will be reproduced, and exchange information will be acquired from ET2026 (steps 21041 and 21042).

[0111] If the status frame from both disk array subset "#0" and "#1" gathers (step 21043), conversion to the main status frame from LU "#0" will be performed after a check of that the status is normal termination (step 21044), and ** status frame elimination will be carried out (step 21045). And IC2023 transmits the command frame for reporting normal termination to a host (the arrow head of drawing 20 (h): step 21046). Finally SP2021 eliminates the exchange information on ET2026 (step 21047).

[0112] The light processing in a mirroring configuration is completed above. Although lead processing to LU "#01" by which mirroring was carried out is performed almost like the light processing which the directions of a data transfer only differ and was mentioned above, unlike a light, it does not need to publish a lead command to two sets of disk array subsets, and should just publish a command frame to either. For example, because of improvement in the speed, although a command frame may always be published to the main LU, if a load is distributed by publishing a command frame by turns etc. to LU of the Lord/******, it is effective.

[0113] In the processing mentioned above, a synchronization of waiting and both is taken for the response of two sets "#0" of disk array subsets, and "#1" at step 21014 and step 21043, and processing is advanced. In such control, after a success of processing by both disk array subsets is checked, in order that processing may progress, the correspondence at the time of error generating becomes easy. On the other hand, in order that the whole processing speed may be dependent on the response of which or the later one, there is a fault that the engine performance falls.

[0114] Since this problem is solved, when it progresses to the next processing, without waiting for the response of a disk array subset or there is a response from one of the disk array subsets, in a disk array switch, it is also possible to carry out "asynchronous" control which progresses to the next processing. In drawing 20 , a broken-line arrow head shows an example of the frame sequence at the time of performing asynchronous control.

[0115] In the frame sequence shown by the broken-line arrow head, transmission of the data transfer preparation-completion frame to the host to whom it is carried out at step 21016 is carried out after processing of step 21009, without waiting for the data transfer preparation-completion frame from the disk array subset 10. In this case, the data transfer preparation-completion frame transmitted to a host is generated by SC2022 of the disk array switch 20 (broken-line arrow head (d')).

[0116] From a host 30, a data frame is transmitted to the disk array switch 20 to the timing shown by the broken-line arrow head (e'). With the disk array switch 20, this data frame is once stored in FB2025. SC2022 answers reception of the data transfer preparation-completion frame from the disk array subset 10, and transmits the data frame held at FB2025 to the disk array subset 10 to which the data transfer preparation-completion frame has been sent (a broken-line arrow head (f0'), (f1')).

[0117] The termination report to a host 30 from the disk array switch 20 is performed when there is a report (a broken-line arrow head (g0'), (g0')) from both disk array subsystems 10 (broken-line arrow head (h')). It is possible only for the part of the time amount Ta shown in drawing 20 to shorten the processing time by such processing.

[0118] The following processings are carried out when an error occurs in the middle of the frame transfer between the disk array switch 20 and the disk array subset 10.

[0119] When the processing under activation is light processing, retry processing is performed to LU which the error generated. If a retry is successful, processing will be continued as it is. When the retry of the count of the convention set up beforehand goes wrong, the disk array switch 20 forbids access to this disk array subset 10 (or LU), and registers into DCT2027 the information which shows that. Moreover, the disk array switch 20 notifies that to the disc system configuration means 70 via MP200 and the communication link controller 204.

[0120] The disc system configuration means 70 answers this notice, and publishes an alarm to an administration terminal 5. Thereby, a manager can recognize that the trouble occurred. Then, the disk array switch 20 continues operation using a normal disk array subset. A host 30 does not recognize that the error occurred and can continue processing.

[0121] raising the failure-proof nature of a disk, since a mirror configuration is realizable with two sets of disk array subsystems according to this operation gestalt — things are made. Moreover, a disk array controller, disk array I/F, and the failure-proof nature of a disk array I/F node can be raised, there is nothing with ** which carries out doubleness of an internal bus etc., and the dependability of the whole disk array system can be raised.

[0122] Three or more sets of the [7th operation gestalt], next the disk array subsets 10 are unified, and how to constitute the group of one set of a logical disk array subset is explained. Data are distributed and stored in two or more disk array subsets 10 with this operation gestalt. Thereby, access to a disk array subset is distributed and a total throughput is raised by inhibiting concentration of access to a specific disk array subset. With this operation gestalt, such striping processing is carried out with a disk array switch.

[0123] Drawing 23 is the address map of the disk array system 1 in this operation gestalt. Striping of the address space of the disk array subset 10 is carried out in the SUTOREIPU size S. The address space of the disk array system 1 seen from the host is distributed by every stripe size S a disk array subset "#0", "#1", "#2", and "#3." Although the size of the stripe size S is arbitrary, the direction which is not not much small is good. When the stripe size S was too small and the stripe crossover to which the data which should be accessed belong to two or more stripes occurs, a possibility that an overhead may occur is in the processing. If stripe size S is enlarged, since the probability for a stripe crossover to occur will decrease, for the improvement in the engine performance, it is desirable. The number of LUs can be set as arbitration.

[0124] Hereafter, actuation of the host I/F node 203 in this operation gestalt is explained and explained paying attention to difference with the 1st operation gestalt, referring to the operation flow chart shown in drawing 24 . In addition, with this operation gestalt, "Striped" is set to CLU Class of the information about the host LU by whom striping was done on the host LU configuration table 20271 of DCT2027, and stripe size "S" is set to CLU Stripe Size.

[0125] If a host 30 publishes a command frame, the disk array switch 20 will recognize that SC2022 which receives this (step 22001) needs to search DCT2027, and needs to carry out striping of this command frame using reception and SP2021 from IC2023 by IC2023 of the host I/F node 203 (step 22005).

[0126] Next, SC2022 searches DCT2027 by SP2021, asks for the stripe number of the stripe with which the data set as the object of access belong from the configuration information containing the stripe size S, and specifies in which disk array subset 10 this stripe is stored (step 22006). Under the present circumstances, although a stripe crossover may occur, about the processing in this case, it mentions later. When a stripe crossover does not occur, based on SP's2021 count result, SC2022 changes to a command frame (step 22007), and stores exchange information in ET2026 (step 22008). Henceforth, the same processing as the 1st operation gestalt is performed.

[0127] When a stripe crossover occurs, SP2021 generates two command frames. This generation is performed with reproducing the command frame which the host 30 published. The frame header of the command frame to generate, a frame payload, etc. are set up newly. Although it is also possible like the 6th operation gestalt to carry out conversion after creating the duplicate of a command frame by SC2022, it shall be newly created by SP2021 here. SC2022 will transmit these to each disk array subset 10, if two command frames are generated.

[0128] Then, data transfer is carried out like the 1st operation gestalt. Here, unlike the 1st operation gestalt or the 6th operation gestalt, it is necessary to transmit the data itself with this operation gestalt in one set of the two sets of a host 30 and the disk array subsets 10. For example, in lead processing, it is necessary to transmit all the data frames transmitted from two sets of the disk array subsets 10 to a host 30. Under the present circumstances, to the data frame transmitted from each disk array subset 10, according to the exchange information registered into ET2026, SC2022 is suitable sequence, adds suitable exchange information and transmits to a host 30.

[0129] In light processing, it transmits to the disk array subset 10 which divides and corresponds to two data frames like the case of a command frame. In addition, the sequence control of a data frame is not indispensable if the host or the disk array subset supports the processing in random order called an AUTOOBU order (Out of Order) function.

[0130] If all data transfer is completed and the disk array switch 20 finally receives two status frames from the disk array subset 10, SP2021 (or SC2022) will create the status frame to a host 30, and will transmit this to a host 30 by IC2023.

[0131] Since access can be distributed to two or more disk array subsets, while being able to raise a throughput as total according to this operation gestalt, also as for an access latency, it is possible to make it decrease on the average.

[0132] Creation of the duplicate for two sets (or disk array subset) of the [8th operation gestalt], next disk array systems is explained as the 8th operation gestalt. A system which is explained here arranges one side of two sets of disk array systems to a remote place, and is equipped with the resistance over the failure of the disk array system of another side by a natural disaster etc. The thing of creation of the duplicate performed between a disaster recovery, and a call and the disk array system of a remote place in a cure to such a disaster is called a remote copy.

[0133] Since mirroring explained with the 6th operation gestalt constitutes a mirror from the disk array subset 10 installed in the geographical almost same location, disk array I/F21 is good by the fiber channel. However, when the disk array (disk array subset) which performs a remote copy is installed in the remote place exceeding 10km, you cannot transmit a frame without junction by the fiber channel. Since [this] the distance between each other is usually set to hundreds of km or more when used for a disaster recovery, it is impossible practically to connect between disk arrays by the fiber channel, and a high-speed public line, satellite communication, etc. by ATM (Asynchronous Transfer Mode) etc. are used.

[0134] Drawing 25 is an example of the disaster recovery structure of a system in this operation gestalt.

[0135] 81 is Site A, 82 is Site B, and both sites are installed in a geographical remote place. 9 is a public line and an ATM packet passes through this. A site A81 and a site B82 have the disk array system 1, respectively. Here a site A81 is a common site usually used, and a site B82 is a remote disaster recovery site used when a site A81 is downed with disaster etc.

[0136] The contents of the disk array subset "#0" of the disk array system 10 of a site A81 and "#1" are copied to the disk array subset for a remote copy of the disk array system 10 of a site B82 "#0", and "#1." What is connected to a remote site among the I/F nodes of the disk array switch 20 is connected to the public line 9 using ATM. This node is called the ATM node 205. The ATM node 205 is constituted like the host I/F node shown in drawing 5 , and IC2023 changes an ATM-fiber channel. This conversion is realized by the same approach as conversion of the SCSI-fiber channel in the 4th operation gestalt.

[0137] Processing of the remote copy in this operation gestalt is similar with processing of mirroring in the 6th operation gestalt. Hereafter, a different point from processing of mirroring in the 6th operation gestalt is explained.

[0138] If a host 30 publishes a light command frame, the disk array system 10 of a site A81 will double a frame like the case in the 6th operation gestalt, and will transmit one of these to the own disk array subset 10. The frame of another side is changed into an ATM packet from a fiber channel frame by the ATM node 205, and is sent to a site B82 through a public line 9.

[0139] To a site B82, the ATM node 205 of the disk array switch 20 receives this packet. IC2023 of the ATM node 205 reproduces a fiber channel frame from an ATM packet, and transmits it to SC2022. SC2022 performs frame conversion like the time of receiving a light command from a host 30, and transmits it to the disk array subset for a remote copy. Henceforth, in data transfer preparation-completion frames, data frames, and all the status frames, a remote copy is realizable by performing fiber channel-ATM conversion in the ATM node 205, and carrying out same frame transfer processing.

[0140] When a host 30 publishes a lead command frame, the disk array switch 20 transmits a command frame only to the disk array subset 10 of a self-site, and leads data only from the disk array subset 10 of a self-site. The actuation at this time becomes the same as that of the 1st operation gestalt.

[0141] According to this operation gestalt, user data can be backed up on real time and it can have the resistance over the site failure by a natural disaster etc., and disk array system failure.

[0142] Integration of two or more LUs included by one set of the [9th operation gestalt], next the disk array subset 10 is explained. For example, in order that the disk unit for main frames may maintain compatibility with the past system, the maximum of the size of a logical volume is set as 2GB. When sharing such a disk array system also with an open system, LU will receive a limit of logical volume size as it is, and many its LUs of small size can be seen from a host. By such approach, when large capacity-ization progresses, the problem that employment becomes difficult arises. Then, it considers unifying this logical volume (namely, LU) and constituting one big integration LU by the function of the disk array switch 20. In this operation gestalt, integration LU is created with the disk array switch 20.

[0143] Integration of LU in this operation gestalt is the same as that of creation of the integration LU by two or more disk array subsets 10 which can be set in the 1st operation gestalt. Difference is only integration by the plurality LU in the same disk array subset 10. The actuation as a disk array system becomes completely the same as that of the 1st operation gestalt.

[0144] Thus, by unifying two or more LUs included by the same disk array subset 10, and creating one big LU, it becomes unnecessary to manage many LUs from a host, and excels in operability, and the disk array system which reduced management cost can be built.

[0145] The setting approach of the [10th operation gestalt], next the shift pass by the disk array switch 10 is explained referring to drawing 26 .

[0146] The configuration of each part in the computing system shown in drawing 26 is the same as that of the 1st operation gestalt. Here, if two sets of hosts 30 access the disk array subset 10 using respectively different disk array I/F21, it will be assumed that it constitutes like. By a diagram, only the number which needs the host I/F node 203 and the disk array I/F node 202 of a disk array subset and the disk array switch 20 for explanation here is shown.

[0147] The disk array subset 10 had the same configuration as drawing 2 , and has connected two disk array I/F controllers to one set of the disk array switch 20, respectively. The shift pass of disk array I/F21 is set to DCT227 of each node of the disk array switch 20. Shift pass is pass of the alternative established so that it may become accessible, also when a failure occurs on one certain pass. Here, the shift pass of disk array I/F "#1" and disk array I/F "#1" is set for the shift pass of disk array I/F "#0" as disk array I/F "#0." Similarly, shift pass is set also about each between the high order adapters in the disk array subset 10, between a cache and alternate memory, and between low order adapters.

[0148] Next, as shown in drawing 26 , disk array I/F21 linked to the high order adapter "#1" of the disk array subset 1 is disconnected, it assumes that the failure occurred, and the setting-operation of shift pass is explained. It becomes impossible for the host "#1" using disk array I/F21 which the failure generated to access the disk array subset 10 at this time. It is recognized as the failure having generated the disk array switch 20 on this pass, when not recovering, even if it detected the abnormalities of the frame transfer between the disk array subsets 10 and carried out retry processing.

[0149] If the failure of pass occurs, SP2021 will register that the failure occurred in disk array I/F "#1" into DCT2027, and will register using disk array I/F "#0" as shift pass. Henceforth, SC2022 of the host I/F node 203 operates so that the frame from a host "#1" may be transmitted to the disk array I/F node 202 linked to disk array I/F "#0."

[0150] The high order adapter 101 of the disk array subset 10 succeeds and processes the command from a host "#1." Moreover, the disk array switch 20 notifies generating of a failure to the disk array system-configuration-control means 70, and generating of a failure is notified to a manager by the disk array system-configuration-control means 70.

[0151] According to this operation gestalt, a change on the shift pass at the time of a failure occurring on pass can be performed without making it recognize to a host side, and a shift processing setup by the side of a host can be made unnecessary. Thereby, the availability of a system can be raised.

[0152] Each operation gestalt explained above explained the disk array system which used the disk unit altogether as storage media. However, this invention is not limited to this, and when not only a disk unit but an optical disk unit, a tape unit, DVD equipment, a semiconductor memory, etc. are used as storage media, it can be applied similarly.

[0153]

[Effect of the Invention] According to this invention, the storage system which can realize easily escape of the storage system according to the scale of a computing system, a demand, etc., improvement in dependability, etc. is realizable.

[Translation done.]

TECHNICAL FIELD

[Field of the Invention] This invention relates to the approach of improvement in the speed of a disk control system, low-cost-izing, and improvement in cost performance especially about the implementation approach of the disk control system which controls two or more disk units.

[Translation done.]

PRIOR ART

[Description of the Prior Art] There is a disk array system which controls two or more disk units as a store system used for a computer system. About a disk array system, it is "A Case for Redundant Arrays of Inexpensive Disks" (RAID), for example.; It is indicated by InProc.ACM SIGMOD and June 1988 (University of California at Berkeley issue). A disk array is operating two or more disk units to juxtaposition, and is the technique of realizing improvement in the speed compared with the storage system which used the disk unit alone.
[0003] There is an approach which used Fabric of a fiber channel (Fibre Channel) as an approach of connecting two or more disk array systems mutually with two or more hosts. the example of the computing system which applied this approach -- Nikkei electronics 1995.7.3 (no.639) "Serial SCSI is to commercial scene still more" P.79 It is shown in drawing 3 . In the computing system indicated here, two or more host computers (below, it is only called a host) and two or more disk array systems are connected to fabric equipment through a fiber channel, respectively. Fabric equipment is the switch of a fiber channel and connects the transfer way between the equipment of the arbitration linked to fabric equipment. Fabric equipment is transparency to a transfer of the "frame" which is the packet of a fiber channel, and a host and a disk array system communicate in two points, without being conscious of fabric equipment of each other.

[Translation done.]

EFFECT OF THE INVENTION

[Effect of the Invention] According to this invention, the storage system which can realize easily escape of the storage system according to the scale of a computing system, a demand, etc., improvement in dependability, etc. is realizable.

[Translation done.]

TECHNICAL PROBLEM

[Problem(s) to be Solved by the Invention] In the conventional disk array system, the number of a disk unit is increased for large-capacity-izing, and if it is going to realize the controller which has the engine performance which balanced the number for high-performance-izing, the performance limit of the internal bus of a controller and the performance limit of the processor which performs transfer control will actualize. In order to cope with such a problem, an internal bus is extended and increasing the number of processors is performed. However, the method of such management causes the complication of control software and the increment in an overhead by complication of a controller configuration, exclusive control of the share data of interprocessor, etc. by much bus control. For this reason, while raising cost very much, the engine performance is reaching the ceiling, consequently cost performance gets worse. Moreover, in a large-scale system, although such equipment can realize the engine performance corresponding to the cost, there is a technical problem which does not balance that the increase of a development cycle and the rise of development cost to which expandability is restricted are caused in the system whose scale is not so large.
[0005] By putting two or more disk array systems in order, and interconnecting with fabric equipment, it is possible to perform large-capacity-izing as the whole system and high performance-ization. However, by this approach, since it cannot be distributed to other equipments even if it is irrelevant between disk array systems and access concentrates on a specific disk array system, high performance-ization on real use is unrealizable. Moreover, since the capacity of the logical disk unit (it is called Logical unit) seen from the host is restricted to the capacity of one set of a disk array system, large capacity-ization of Logical unit is unrealizable.
[0006] Although the mirror configuration by two sets of disk array systems can be realized using the mirroring function which the host has when it is going to form the whole disk array system into high reliance, the control overhead for mirroring by the host occurs, and the technical problem that system performance is restricted occurs. Moreover, if many disk array systems exist according to an individual in a system, a load for a system administrator to manage will increase. For this reason, management cost increases -- many maintenance staffs and the maintenance costs for two or more sets are needed. Furthermore, since two or more disk array systems and fabric equipment are isolated systems, respectively, it is necessary to carry out various setup by different approach for every equipment. For this reason, employment cost increases with a manager's training and increase of an operate time.
[0007] The purpose of this invention solves the technical problem in these conventional technique, can build the storage system according to the scale of a computing system, a demand, etc., and is to realize the storage system which can respond to the escape of the storage system in the future, improvement in dependability, etc. easily.

[Translation done.]

MEANS

[Means for Solving the Problem] The store which has the storage with which the store system of this invention holds data, Two or more store subsystems which have the control device which controls this store, The 1st interface node connected to the computer which uses the data held at two or more store subsystems, Two or more 2nd interface nodes by which each was connected to either of the store subsystems, And the 1st interface node and two or more 2nd interface nodes are connected, and it has a transfer means to transmit a frame between the 1st interface node and two or more 2nd interface nodes.
[0009] Preferably, the 1st interface node answers the frame sent from a calculating machine, analyzes this frame, carries out signal transduction to the configuration management table which stored store structure-of-a-system information about the destination of the frame based on the configuration information held at the configuration management table, and is transmitted to a transfer means.
[0010] Moreover, on the occasion of a transfer of a frame, the 1st interface node adds the node address information on the node which should receive the frame to a frame. A transfer means transmits a frame according to the node address information added to the frame. The 2nd interface node carries out the reconstititution of the frame except for node address information from the frame received from the transfer means, and transmits it to the target store subsystem.
[0011] In a mode with this invention, a storage system has a management processor linked to a transfer means. A management processor sets configuration information as a configuration management table according to the directions from an operator. The information which restricts access from a computer is included in configuration information.
[0012]
[Embodiment of the Invention] [1st operation gestalt] drawing 1 is a block diagram in 1 operation gestalt of the computer system using the disk array system by which this invention was applied.
[0013] It is the host computer (host) to which, as for 1, a disk array system is connected to, and, as for 30, a disk array system is connected. The disk array system 1 has the communication interface 80 between the disk array system-configuration-control means 70, and the disk array switch 20 and the disk array system-configuration-control means 70 of performing setting management of the disk array subset 10, the disk array switch 20, and the whole disk array system, and between the disk array subset 10 disk-array system-configuration-control means 70 (communication link I/F). It connects with the host interface (host I/F) 31, and a host 30 and the disk array system 1 connect host I/F31 to the disk array switch 20 of the disk array system 1. In the interior of the disk array system 1, the disk array switch 20 and the disk array subset 10 are connected with a disk array interface (disk array I/F21).
[0014] Although a host 30 and four disk array subsets 10 are shown respectively, they are [no limit] about this number and are arbitrary by a diagram. The number of a host 30 and the disk array subset 10 may differ. Moreover, the disk array switch 20 is doubled with this operation gestalt as illustration. Each host 30 and each disk array subset 10 are connected to the both sides of the disk array switch 20 doubled by respectively separate host I/F31 and disk array I/F21. This is for enabling access to the disk array system 1 from a host 30 by using another side, even if one disk array switch 20, host I/F31, or disk array I/F21 breaks down, and realizing high availability. However, such doubleness is not necessarily indispensable and is selectable according to the reliability level required of a system.
[0015] Drawing 2 is the block diagram showing the example of 1 configuration of the disk array subset 10. The high order adapter which 101 interprets the command from host system (host 10), carries out a cache hit mistake judging, and controls the data transfer between host system and a cache, the shared memory (it is called a cache and a shared memory below) in which 102 stores the cache for disk data-access improvement in the speed and the share data between multiprocessors, and 104 are two or more disk units stored in the disk array subset 10. 103 is a low order adapter which controls a disk unit 104 and controls the data transfer between a disk unit 104 and a cache. 106 is a disk array subset configuration management means, communicates through the disk array system-configuration-control means 70 and

communication link I/F80 which manage the disk array system 1 whole, and manages a setup of a configuration parameter, the report of fault information, etc.

[0016] The high order adapter 101, a cache and a shared memory 102, and the low order adapter 103 are doubled, respectively. Like doubleness of the above-mentioned disk array switch 20, this reason is for realizing the sex for Takayoshi, and is not indispensable. Moreover, either of the doubled low order adapters 103 of each disk unit 104 is also controllable. Although the same memory means is shared from a viewpoint of low-cost-izing to the cache and the shared memory with this operation gestalt, of course, these can also be dissociated.

[0017] The high order adapter 101 includes the high order bus 1012 which performs communication link between a cache and a shared memory 102, and a high order MPU 1010 and the disk array I/F controller 1011, and data transfer. [the high order MPU 1010 which performs control of the high order adapter 101, host system 1011, i.e., the disk array I/F controller which controls disk array I/F21 which is connection I/F with the disk array switch 20, and]

[0018] Although one disk array I/F controller 1011 is shown every high order adapter 101 by a diagram, two or more disk array I/F controllers 1011 may be formed to one high order adapter.

[0019] The low order adapter 103 includes the low order bus 1032 which performs communication link between a cache and a shared memory 102, and the low order MPU 1030 and the disk I/F controller 1031 which control disk I/F which is an interface with the low order MPU 1030 which performs control of the low order adapter 103, and a disk 104, and data transfer. [the disk I/F controller 1031, and]

[0020] Although four disk I/F controllers 1031 are shown every low order adapter 103 by a diagram, the number is arbitrary and can be changed according to the configuration and the number of a disk to connect of a disk array.

[0021] Drawing 3 is the block diagram showing the example of 1 configuration of the disk array switch 20. The management processor (MP) which is a processor to which 200 performs control and management of the whole disk array switch, the crossbar switch with which 201 constitutes the mutual switch path of nxn, the disk array I/F node in which 202 is prepared every disk array I/F21, the host I/F node in which 203 is prepared every host I/F31, and 204 are communication link controllers which perform the communication link between the disk array system-configuration-control means 70. I/F between clusters for the pass whose 2020 connects a crossbar switch 201 with the disk array I/F node 202, the pass on which 2030 connects a crossbar switch 201 with the host I/F node 203, and 2040 to connect with other disk array switches 20, and constitute a cluster, and 2050 are the pass for connecting a crossbar switch 201 with MP200.

[0022] Drawing 4 is the block diagram showing the structure of a crossbar switch 201. 2010 is a switching port (SWP) which are the pass 2020, 2030, and 2050 linked to a crossbar switch 201, and a port which connects I/F2040 between clusters. SWP2010 has the same structure altogether and performs switching control of the transfer path to other SWP(s) [SWP / a certain]. Although the transfer path is shown only about one SWP by a diagram, the same transfer path exists among all SWP(s).

[0023] Drawing 5 is the block diagram showing the example of 1 configuration of the host I/F node 203. With this operation gestalt, in order to explain concretely, it is assumed that it is what uses a fiber channel for both host I/F31 and disk array I/F21. Of course, it is also possible as host I/F31 and disk array I/F21 to apply interfaces other than a fiber channel. By using the same interface for both the host I/F node 203 and the disk array I/F node 202, both are made to the same structure. In this operation gestalt, it is constituted like the host I/F node 203 which also shows the disk array I/F node 202 in drawing. Below, the host I/F node 203 is explained to an example.

[0024] The retrieval processor which searches to which node 2021 transmits the received fiber channel frame (it is only called a frame below) (SP), 2022 is a host 30 (in the case of the disk array I/F node 202). The interface controller which transmits and receives a frame between the disk array subsets 10 (IC), The switching controller which changes based on the result with which SP2021 searched 2022 to the frame which IC2023 received (SC), The packet generation section packet-ized in the format that a crossbar switch 201 can be passed in order that 2024 may transmit the frame which SC2021 changed to other nodes (SPG), The frame buffer which stores temporarily the frame which 2025 received (FB), The exchange table which manages an exchange number for 2026 to identify the exchange (Exchange) which are two or more frame trains which corresponded to the disk array access request command (it is only called a command below) from one host (ET), 2027 is a disk array configuration management table (DCT) which stores the configuration information of two or more disk array subsets 10.

[0025] As for each configuration sections of all of the disk array switch 20, it is desirable on the engine performance to consist of hardware logic. However, if the engine performance called for can be satisfied, it is also possible to realize the function of SP2021 or SC2022 by the program control using a general-purpose processor.

[0026] Each disk array subset 10 has managed the disk unit 104 which each has as 1 or two or more logical disk units. This logical disk unit is called Logical unit (LU). LU does not need to correspond by the physical disk unit 104 and 1 to 1, and two or more LUs may be constituted by one set of a disk unit 104, or one LU may consist of two or more disk units 104.

[0027] When it sees from the outside of the disk array subset 10, one LU is recognized as one set of a disk unit. With this operation gestalt, still more logical LU is constituted by the disk array switch 20, and a host 30 operates so that it may access to this LU. On these specifications, when one LU recognized from a host 30 consists of an independent LU (ILU) and two or more LUs in LU recognized by the host 30 when one LU recognized from a host 30 consists of one LU, LU recognized by the host 30 is called Integration LU (CLU).

[0028] The correspondence relation of the address space between each hierarchy in case one integration LU is constituted from an LU of four disk array subsets by drawing 12 is shown. In drawing, the address space in one integration LU of the disk array system 1 which saw 1000 from the host "#2", and 1100 show the address space of LU of the disk array subset 10 as an example, and 1200 shows the address space of a disk unit 104 (here, illustrated only about the disk array subset "#0").

[0029] LU of each disk array subset 10 shall be constituted as a RAID5 (Redundant Arrays of Inexpensive Disks Level 5) mold disk array by four sets of disk units 104 here. Each disk array subset 10 has LU which has the capacity of n0, n1, n2, and n3, respectively. The disk array switch 20 unifies the address space which these four LUs have to the address space which has the capacity of (n0+n1+n2+n3), and integration LU recognized from a host 30 is realized.

[0030] With this operation gestalt, when host #2 access a field A1001, the access request which specified the field A1001 is changed into the demand for accessing the field A'1101 of LU of disk array subset #0 with the disk array switch 20, and is transmitted to disk array subset #0, for example. Disk array subset #0 accesses a field A'1101 further by mapping in 1201 A" of fields on a disk unit 104. Mapping between an address space 1000 and an address space 1100 is performed based on the configuration information held at DCT207 which the disk array switch 20 has. About the detail of this processing, it mentions later. In addition, it is the technique already well known about mapping in a disk array subset, and omits about detailed explanation on these specifications.

[0031] In this operation gestalt, DCT207 contains a system configuration table and a subset configuration table. Drawing 6 shows the configuration of a system configuration table, and drawing 7 shows the configuration of a subset configuration table.

[0032] As shown in drawing 7 , the system configuration table 20270 has the host LU configuration disk array I/F node configuration table 20271 and 20272 holding the information which shows Host's LU configuration showing the connection relation between the disk array I/F node 202 of the disk array switch 20, and the disk array subset 10.

[0033] The host LU configuration table 20271 has LU information (LU Info.) which is the information about LU of Host-LU No. which is the number which was seen from the host 30, and which identifies the LU for every LU, LU Type which shows the attribute of LU, CLU Class and CLU Stripe Size, Condition that is the information which shows Host's LU condition, and the disk array subset 10 which constitutes Host LU.

[0034] LU Type is information which shows the class of LU whether this host LU is CLU or to be ILU. CLU Class is information which shows any of "Joined", "mirrored", and "Striped" that class is, when it is shown by LU Type that this host LU is CLU. "Joined" shows that CLU which connects some LUs and has one big storage space is constituted, as drawing 11 explained. "Mirrored" shows that it is LU doubled by two LUs so that it may mention later as the 6th operation gestalt. "Striped" consists of two or more LUs, and shows that it is LU by which data were distributed and stored in LU of these plurality so that it may mention later as the 7th operation gestalt. CLU Stripe Size shows striping size (size of the block used as the unit of distribution of data), when it is shown by CLU Class that it is "Striped."

[0035] There are four kinds of the conditions by which it is shown by Condition, "Normal", "Warning", "Fault", and "Not Defined." As for "Normal", this host LU shows that it is in a normal condition. "Warning" shows that degeneration operation is performed to one corresponding to LU which constitutes this host LU of disk units for the reason of the failure having occurred. "Fault" shows that this host LU cannot be operated by failure of the disk array subset 10 etc. "Not Defined" shows that the corresponding host LU of Host-LU No. is not defined.

[0036] LU Info includes LUN within the information which specifies the disk array subset 10 to which that LU belongs about LU which constitutes this host LU, and a disk array subset, and the information which shows that size. When Host LU is ILU, the information about the only LU is registered. When Host LU is CLU, the information about each LU is registered about all LUs that constitute it. For example, in drawing, it is CLU which consists of four LUs, LUN "0" of a disk array subset "#0", LUN "0" of a disk array subset "#1", LUN "0" of a disk array subset "#2", and LUN "0" of a disk array subset "#3", and, as for Host-LU whose Host-LU No. is "0", it turns out that it is CLU the CLU class of whose is "Joined."

[0037] The disk array I/F node configuration table 20272 holds the information which shows of which disk array switch 20 the disk array I/F node 202 is connected for every port of the disk array subset 10 which disk array I/F21 connects.

[0038] Specifically, it has Subset No. which specifies the disk array subset 10, Subset Port No. which pinpoints a port, Switch No. which specifies the disk array switch 20 linked to the port, and I/F Node No. which specifies the disk array I/F node 202 of the disk array switch 20. When the disk array subset 10 is equipped with two or more ports, information is set up for every port of the.

[0039] A subset configuration table has two or more tables 202720-202723 corresponding to each disk array subset 10, as shown in drawing 7 . Each table contains LU configuration table 202740 holding the information which indicates the configuration of LU built in the disk array subset 10 to be the RAID group configuration table 202730 holding the information which shows a RAID group's configuration built within the disk array subset 10.

[0040] In a configuration of that striping of the RAID level 0 and the 5 grades was carried out, Group No. which shows the number by which the RAID group configuration table 202730 was added to the RAID group, Level which shows the RAID group's level, Disks which is the information which shows the number of the disks which constitute the RAID group, and its RAID group contain as information Stripe Size which shows the stripe size. For example, in the table shown in drawing, a RAID group "0" is a RAID group constituted by four sets of disk units, RAID level is 5 and stripe size is S0.

[0041] LU configuration table 202740 contains as information LU No. which shows the number (LUN) added to LU, RAID Group which shows whether this LU is constituted by which RAID group, Condition which shows the condition of LU, Size which shows the size (capacity) of this LU, Port which shows whether this LU is accessible from the port of disk array subset 10 throat, and Alt.Port which shows the port used as that alternative. The condition by which it is shown by Condition has four kinds, "Normal", "Warning", "Fault", and "Not Defined", like Condition about Host LU. When a failure occurs in the port pinpointed for the information set as Port, it is used, but the port pinpointed using the information set as Alt.Port can also be used in order to only access the same LU from two or more ports.

[0042] Drawing 8 is the block diagram of the frame in a fiber channel. The frame 40 of a fiber channel contains CRC (Cyclic RedundancyCheck)403 which is an error detection code with a frame payload of 402 or 32 bits which is the part which stores SOF (Start Of Frame)400, the frame header 401, and the actual condition data of a transfer in which the head of a frame is shown, and EOF (End Of Frame)404 which shows the tail end of a frame. The frame header 401 has structure as shown in drawing 9 , and ID (SEQ_ID) of a sequence which specifies the frame group under each exchange ID specified by a response place (OX_ID, RX_ID) and exchange is stored the ID [of the frame source] (S_ID), ID [of the frame destination] (D_ID), and starting origin of an exchange.

[0043] With this operation gestalt, ID by which ID assigned to the host 30 as S_ID was assigned as D_ID to the port of the disk array switch 20 again is used for the frame published by the host 30. The exchange ID of one pair (OX_ID, RX_ID) is assigned to one host command. When it is necessary to publish two or more data frames to the same exchange, the same SEQ_ID is assigned to all the data frames, and each is identified at a sequence count (SEQ_CNT). The maximum length of the frame payload 402 is 2110 bytes, and the contents stored for every frame kind differ. For example, in the case of the FCP_CMD frame mentioned later, as shown in drawing 10 , SCSI Logical Unit Number (LUN), Command Description Block (CDB), etc. are stored. CDB contains a command cutting tool required for disk (disk array) access, the transfer initiation logical address (LBA), and transfer length (LEN).

[0044] Hereafter, actuation of the disk array system of this operation gestalt is explained.

[0045] Before using a disk array system, it is necessary to the disk array switch 20 to set up the configuration information of the disk array subset 10. A system administrator acquires all the disk array subsets 10 and the configuration setting information on the disk array switch 20 from an administration terminal 5 through the disk array system configuration means 70. A manager inputs setting information required for various setup, such as a configuration setup of Logical unit, a setup of RAID level, and a setup of the shift pass at the time of failure generating, so that it may become a desired system configuration from an administration terminal 5. The disk array system-configuration-control means 70 receives the setting information, and transmits setting information to each disk array subset 10 and the disk array switch 20. In addition, the 5th operation gestalt explains separately the input of the setting information in an administration terminal 5.

[0046] With the disk array switch 20, the communication link controller 204 acquires setting information, and configuration information, such as address space information on each disk array subset 10, is set up by MP200. MP200 distributes the configuration information of the disk array subset 10 to each host I/F node 203 and the disk array I/F node 202 by crossbar switch 201 course.

[0047] Each nodes 203 and 202 store configuration information in DCT2027 by SP2021, if this information is received. In the disk array subset 10, the disk array subset configuration management means 106 acquires setting information, and stores in a shared memory 102. Each high order MPU 1010 and low order MPU 1030 carry out each configuration management with reference to the setting information on a shared memory 102.

[0048] Below, actuation when a host "#2" publishes a lead command to the disk array system 1 is explained. The flow chart of actuation [in / for the mimetic diagram showing the sequence of the frame transmitted to drawing 11 through a fiber channel at the time of the lead actuation from a host / the host I/F node 203 of the disk array switch at this time] is shown in drawing 13 .

[0049] In addition, in the following explanation, a host "#2" assumes accessing the storage region A1001 in drawing 12 . A " of actual storage regions corresponding to a storage region A1001 shall exist in the address space of disk unit #2 which constitute LU of LUN=0 of a disk array subset "#0." Moreover, "CLU" shall be set to LU Type of the host LU configuration table 20271 which defines LU which constitutes an address space 1000, and "Joined" shall be set to CLU Class.

[0050] A host 30 publishes the command frame "FCP_CMD" which stored the lead command on the disk array switch 20 at the time of the lead of data (drawing 11 arrow head (a)). The host I/F node "#2" of the disk array switch 20 receives a command frame "FCP_CMD" by host I/F31 course by IC2023 (step 20001). IC2023 transmits a command frame to SC2022. SC2022 once stores the received command frame in FB2025. Under the present circumstances, SC2022 calculates CRC of a command frame and inspects that receipt information is right. If an error is in inspection of CRC, SC2022 will notify that to IC2023. IC2023 will report a CRC error to a host 30 through host I/F31, if the notice of an error is received from SC2022. (Step 20002) .

[0051] The frame to which CRC held SC2022 to FB2025 the right case is led, it recognizes that it is a command frame, and the frame header 401 is analyzed (step 20003). And SC2022 is directed to SP2021 and registers exchange information, such as S_ID, D_ID, and OX_ID, into ET2026 (step 20004).

[0052] Next, SC2022 analyzes the frame payload 402 and acquires LUN and CDB which were specified by the host 30 (step 20005). With directions of SC2022, SP2021 searches DCT2027 and gets the configuration information of the disk array subset 10. Specifically, SP2021 finds the information which has Host-LU No. which is in agreement with LUN stored in the frame payload 402 which searched the host LU configuration table 20271 and was received. SP2021 recognizes Host's LU configuration from the information set as LU Type and CLU Class, and distinguishes LUN of the disk subset 10 which should be accessed based on the information currently held at LU Info., and LU in it, and LBA within this LU. Next, with reference to LU configuration table 202740 of the subset configuration table 202720, SP2021 checks the connection port of the target disk array subset 10, and gets node No. of the disk array I/F node 202 linked to the port from the disk array I/F node configuration table 20272. SP2021 reports the conversion information that the disk array subset 10 which carried out in this way and was obtained is identified, such as a number, and LUN, LBA, to SC2022. (Step 20006) .

[0053] Next, SC2022 changes LBA in LUN and CDB of the frame payload 402 using the acquired conversion information. Moreover, D_ID of the frame header 401 is changed into D_ID of the corresponding host I/F controller 1011 of the disk array subset 10. In addition, S_ID is not rewritten at this time (step 20007).

[0054] SC2022 transmits the disk array I/F node number linked to the command frame and the object disk array subset 10 after conversion to SPG2024. SPG2024 generates the packet which added the easy extended header 601 as shown in drawing 14 to the command frame after the received conversion. This packet is called the switching packet (S Packet) 60. S In the extended header 601 of Packet60, a source (self-node) number, a destination node number, and transfer length are ****** rare **. SPG2024 transmits generated S Packet60 to a crossbar switch 201 (step 20008).

[0055] A crossbar switch 201 receives S Packet60 by SWP2010 linked to a host I/F node "#2." SWP2010 performs switch control to SWP which the node of the destination connects with reference to the extended header 601 of S Packet60, establishes a path, and transmits S Packet60 to the disk array I/F node 202 (here disk array I/F node "#0") of the destination. SWP2010 will release the path, if it carries out at every reception of establishment of a path of S Packet60 and a transfer of S Packet60 is completed. In a disk array I/F node "#0", SPG2024 receives S Packet60, removes the extended header 601, and passes the part of a command frame to SC2022.

[0056] SC2022 writes its ID in S_ID of the frame header of the received command frame. Next, to SP2021, it points to SC2022 so that exchange information, such as S_ID of a command frame, D_ID, and OX_ID, and a frame source host I/F node number may be registered into ET2026, and it transmits a command frame to IC2023. IC2023 transmits a command frame to the disk array subset 10 (here disk array subset "#0") to connect according to the information on the frame header 401 (drawing 11 arrow head (b)).

[0057] A disk array subset "#0" receives the command frame after conversion "FCP_CMD" by the disk array I/F controller 1011. It is recognized as a high order MPU 1010 being a command which acquires LUN and CDB which were stored in the frame payload 402 of a command frame, and leads the data of LEN length from LBA of the specified Logical unit.

[0058] A high order MPU 1010 performs cache hit mistake / hit judging with reference to the cache management information stored in the shared memory 102. If it hits, data transfer will be carried out from a çache 102. Since it is necessary to lead data from a disk unit in a mistake, address translation based on the configuration of RAID5 is carried out, and cache space is secured. And processing information required for lead processing from a disk unit 2 is generated, and processing information is stored in a shared memory 102 in order to take over processing to low order MPU 1030.

[0059] Low order MPU 1030 starts processing ignited by processing information having been stored in the shared memory 102. Low order MPU 1030 specifies the suitable disk I/F controller 1031, generates the lead command to a disk unit 2, and publishes a command for the disk I/F controller 1031. The disk I/F controller 1031 stores the data led from the disk unit 2 in the address with which the cache 102 was specified, and notifies a termination report to low order MPU 1030. Low order MPU 1030 stores processing termination information in a shared memory 102 that it should notify that processing was completed correctly to a high order MPU 1010.

[0060] A high order MPU 1010 resumes processing ignited by processing termination information having been stored in the shared memory 102, and notifies lead data-preparation completion to the disk array I/F controller 1011. The disk array I/F controller 1011 publishes "FCP_XFER_RDY" which is a data transfer preparation-completion frame in a fiber channel to the disk array I/F node "#0" concerned of the disk array switch 20 (drawing 11 arrow head (c)).

[0061] In a disk array I/F node "#0", if a data transfer preparation-completion frame "FCP_XFER_RDY" is received, SC2022 will gain the response place exchange ID (RX_ID) which received from the disk array subset 20, will specify S_ID, D_ID, and OX_ID, will direct to SP2021, and will register RX_ID into the exchange information concerned on ET2026. SC2022 gains the host I/F node number of the destination (source of a command frame) of a data transfer preparation-completion frame. SC2022 cancels S_ID of this frame and transmits it to SPG2024. As SPG2024 was described previously, it generates S Packet, and it transmits it to a candidate host I/F node "#2" by crossbar switch 201 course.

[0062] In a host I/F node "#2", if SPG2024 receives S Packet of a data transfer preparation-completion frame, the extended header of S Packet will be removed, "FCP_XFER_RDY" will be reproduced, and SC2022 will be passed (step 20011). SC2022 is directed to SP2021 and specifies the exchange which searchs ET2026 and corresponds (step 20012).

[0063] Next, SC2022 investigates whether a frame is "FCP_XFER_RDY" (step 20013), and if it is "FCP_XFER_EDY", it will direct renewal of the response place exchange ID of ET2026 (RX_ID) to SP2021. The value added to this frame is used as response place exchange ID (step 20014). And SC2022 changes S_ID of the frame header 401, and D_ID into the suitable value which used ID of the host I/F node 203, and a host's 30 ID (step 20015). The frame header 401 is changed into the frame to a host "#2" by these processings. IC2023 publishes this data transfer preparation-completion frame "FCP_XFER_RDY" to a host "#2" (the arrow head of drawing 11 (d): step 20016).

[0064] In order that the disk array I/F controller 1011 of a disk array subset "#0" may perform data transfer, it generates a data frame "FCP_DATA" and transmits it to the disk array switch 20 (drawing 11 arrow head (e)). Since the transfer length of a frame payload has a limit, the greatest data length which can be transmitted by one frame is 2KB. When a data length exceeds this, only a required number generates and publishes a data frame. The same SEQ_ID is assigned to all data frames. Issue of a data frame is the same as that of the case where it is a data transfer preparation-completion frame except for two or more frames being generated to the same SEQ_ID (that is, SEQ_CNT changing).

[0065] The disk array switch 20 changes the frame header 401 of a data frame "FCP_DATA" like processing of a data transfer preparation-completion frame. However, since RX_ID is already established in the transfer of a data frame, processing of step 20014 in processing of a data transfer preparation-completion frame is skipped. The disk array switch 20 transmits a data frame to a host "#2" after conversion of the frame header 401 (drawing 11 arrow head (f)).

[0066] Next, in order that the disk array I/F controller 1011 of a disk array subset "#0" may perform an exit-status transfer, it generates a status frame "FCP_RSP" and publishes it to the disk array switch 20 (drawing 11 arrow head (g)). With the disk array switch 20, like processing of a data transfer preparation-completion frame, SPG2024 removes an extended header from S Packet, reproduces a "FCP_RSP" status frame (step 20021), searches ET2026 by SP2021, and acquires exchange information (step 20022). SC2022 changes a frame based on the information (20023 of a step). The changed frame is transmitted to a host "#2" by IC2023 (drawing 11 arrow head (h): step 20024). Finally SP2021 deletes exchange information from ET2026 (step 20025).

[0067] Lead processing from a disk array is performed as mentioned above. Lead processing mentioned above only by the direction of transfer of a data frame being reversed also about the light processing to the disk array system 1 and same processing are performed.

[0068] As shown in drawing 3 , the disk array switch 20 equips the crossbar switch 201 with I/F2040 between clusters. I/F2040 between clusters is not used in the system configuration shown in drawing 1 . The disk array switch 20 of this operation gestalt is mutually connectable with other disk array switches, as shown in drawing 15 using I/F2040 between clusters.

[0069] the disk array switch 20 in this operation gestalt — although the disk array subset 10 is connectable only to a total of eight sets with a host 30 if independent, the number of disk arrays can be increased with the host 10 who interconnects and can connect two or more disk array switches using I/F2040 between clusters. For example, in the system shown in drawing 15 , the disk array subset 10 can be connected to a total of 32 sets with a host 30 using four sets of the disk array switches 20, and data transfer becomes possible mutually among these.

[0070] Thus, with this operation gestalt, the connection number of a disk array subset or a host can be increased according to the need for disk capacity or the engine performance. Moreover, since between host-disk array systems is connectable using host I/F for a required transfer band, the expandability of capacity, the engine performance, and the number of connection can be raised sharply.

[0071] According to the operation gestalt explained above, even if the engine performance of one set of a disk array subset is restricted by internal MPU and an internal internal bus, it can interconnect between disk array subsets with a host with a disk array switch using two or more disk array subsets. Thereby, the engine performance high as a disk array system total is realizable. Even if the engine performance of a disk array subset is comparatively low, high performance-ization is realizable by using two or more disk array subsets. Therefore, according to the scale of a computer system, only the required number can connect the disk array subset of low cost, and it becomes possible to build a disk array system at the suitable cost according to a scale.

[0072] Moreover, what is necessary is just to add a disk array subset as required when increase of disk capacity and improvement in the engine performance are needed. Furthermore, since the host and disk array subset of a number of arbitration are connectable using two or more disk array switches, both capacity the engine performance and the number of connection can be raised sharply, and the system which has high expandability can be realized.

[0073] Since the contraction machine of the conventional disk array system itself can be used as a disk array subset according to this operation gestalt, the already developed large-scale control-software property can be used as it is, and reduction of development cost and compaction of a development cycle can be realized further again.

[0074] [2nd operation gestalt] drawing 16 is the block diagram of the computer system in the 2nd operation gestalt of this invention. This operation gestalt changes only the frame header 401 in the host I/F node of a disk array switch, and the frame payload 402 is different from the 1st operation gestalt constitutionally at the point which the point which is not operated and a disk array switch, host I/F, and disk array I/F have not doubled. Therefore, the configuration of each part does not have the 1st operation gestalt and the place which changes a lot, and omits explanation about the detail.

[0075] In drawing 16 , each disk array subset 10 consists of two or more Logical unit (LU) 110. Each LU110 is constituted as an independent LU. Generally, LUN assigned to LU110 in each disk array subset 10 is the consecutive number which begins from 0. For this reason, to show continuously LUN of all LUs110 in the disk array system 1 to a host 30, it is necessary to change the LUN field of the frame payload 402 like the 1st operation gestalt. With this operation gestalt, by showing a host 30 LUN of each disk array subset 10 as it is, conversion of the frame payload 402 is made unnecessary and control of a disk array switch is made easy.

[0076] The disk array switch 20 of this operation gestalt is assumed to be what can access the specific disk array subset 10 every host I/F node 203. In this case, if one host I/F31 is used, only LU110 in one set of the disk array subset 10 is accessible. The host is connected to two or more host I/F nodes 203 to access LU110 of two or more disk array subsets 10 from one set of a host. Moreover, when enabling it to access LU110 of one set of the disk array subset 10 from two or more hosts 30, loop-formation topology, fabric topology, etc. are used for the same host I/F node 203, and two or more hosts 30 are connected. Thus, since the disk array subset 10 will be decided for every D_ID of the host I/F node 203 in case one LU110 is accessed from one set of a host 30 if constituted, it is possible to show a host 30 LUN of each LU as it is.

[0077] With this operation gestalt, for the reason mentioned above, since LUN of LU110 in each disk array subset 10 is shown to the host 30 as it is at the host 30, conversion of LUN in the disk array switch 20 becomes unnecessary. For this reason, the disk array switch 20 will change only the frame header 401 like the 1st example, if a frame is received from a host 30, and it transmits the frame payload 402 to the disk array subset 10, without changing. Since actuation of each part in this operation gestalt is the same as that of the 1st operation gestalt if it removes that conversion of the frame payload 402 is not performed, detailed explanation is omitted here. According to this operation gestalt, development of the disk array switch 20 can be made easy.

[0078] With the 2nd operation gestalt of the [3rd operation gestalt], in the host I/F node of a disk array switch, although only the frame header is changed, by the 3rd operation gestalt explained below, the gestalt which does not change a frame is explained also including a frame header. The computer system of this operation gestalt is constituted like the computer system in the 1st operation gestalt shown in drawing 1 .

[0079] With the 1st and 2nd operation gestalt, internal configurations of the disk array system 1, such as the number of the disk array subset 10 and a configuration of LU110, are concealed to a host 30. For this reason, from a host 30, the disk array system 1 appears as one storage on the whole. On the other hand, the disk array subset 10 is opened to a host 30 as it is, and a host 30 enables it to use ID of the port of a direct disk array subset as D_ID of a frame header with this operation gestalt. Thereby, a disk array switch just needs to control a transfer of a frame according to the information on a frame header, and can change and use switching equipment equivalent to the fabric equipment of the fiber channel in the conventional technique for the disk array switch 20.

[0080] The disk array system-configuration-control means 70 communicates with the means of communications 204 of the communication link controller 106 of the disk array subset 10, and the disk array switch 20, and gains or sets up the configuration information of each disk array subset 10 and the disk array switch 20.

[0081] The disk array switch 20 has fundamentally the same configuration as the disk array switch in the 1st operation gestalt shown in drawing 3 . However, in order to control a transfer of a frame by this operation gestalt, using the information on the frame header of the frame which a host 30 publishes as it is, the function of conversion, such as DCT2027 which the host I/F node 203 of the disk array switch 20 and the disk array I/F node 202 have with the 1st operation gestalt or the 2nd operation gestalt, and a frame header realized by SC2022 and SPG2024 grade, becomes unnecessary. The crossbar switch 201 which the disk array switch 20 has transmits the frame of a fiber channel according to the information on a frame header between the host I/F node 203 and the disk array I/F node 202.

[0082] With this operation gestalt, in order to manage the disk array structure of a system collectively with the disk array system-configuration-control means 70, the disk array system-configuration-control means 70 is equipped with a disk array administrative table (this table is also hereafter called DCT). DCT with which the disk array system-configuration-control means 70 is equipped contains two table groups, the system configuration table 20270 shown in drawing 6 and 7, and the subset configuration tables 202720-202723. In addition, with this operation gestalt, since Host LU is altogether constituted as ILU, all LU Type (s) of the host LU configuration table 20271 serve as "ILU", and CLU Class and CLU Stripe Size do not make semantics.

[0083] A manager operates an administration terminal 5, communicates with the disk array system-configuration-control means 70, acquires information, such as disk capacity of the disk array subset 10, and the number of a disk unit, and performs a setup of LU110 of the disk array subset 10, a setup of RAID level, etc. Next, a manager communicates with the disk array system-configuration-control means 70 with an administration terminal 5, controls the disk array switch 20, and sets up the related information between the disk array subsets 20 with each host 30.

[0084] LU110 can come to be seen as the configuration of the disk array system 1 is established and a manager wishes from a host 30 by the above actuation. The disk array configuration management means 70 can save the above setting information, and according to the actuation from a manager, the check of a configuration is performed and it can make a change of a configuration.

[0085] Once it constitutes the disk array system 1 according to this operation gestalt, existence of the disk array switch 20 cannot be made to be able to recognize from a manager, and two or more disk array subsystems can be treated like one set of a disk array system. Moreover, according to this operation gestalt, the disk array switch 20 and the disk array subset 10 can be systematically operated according to the same operating environment, and the configuration check and a configuration change also become easy. Furthermore, without changing a setup of a host 30, when transposing the disk array system which was being used conventionally to the disk array system in this operation gestalt according to this operation gestalt, it can double with the disk array structure of a system which was using the configuration of the disk array system 1 till then, and compatibility can be maintained.

[0086] With the 1st to 3rd operation gestalt explained beyond the [4th operation gestalt], the fiber channel is used for host I/F. The operation gestalt explained below explains the gestalt in which interfaces other

than a fiber channel were intermingled.
[0087] Drawing 17 shows the example of 1 configuration of IC2023 of the host I/F node 203 interior in case host I/F is Parallel SCSI. The SCSI protocol controller (SPC) by which 20230 performs protocol control of Parallel SCSI, the fiber channel protocol controller (FPC) by which 20233 performs protocol control of a fiber channel, the protocol conversion processor (PEP) to which 20231 carries out protocol conversion of Serial SCSI of a fiber channel to Parallel SCSI, and 20232 are buffers (BUF) which save the data in protocol conversion temporarily.
[0088] In this operation gestalt, a host 30 publishes the SCSI command to the disk array I/F node 203. In the case of a lead command, SPC20230 stores this in BUF20232, and reports reception of a command to PEP20231 by interruption. PEP20231 uses the command stored in BUF20232, changes it into the command to FPC20233, and is sent to FPC20233. If this command is received, FPC20233 will be changed into a frame format and will be handed over to SC2022. Under the present circumstances, Exchange ID, Sequence ID, Source ID, and Destination ID are added by PEP20231 so that subsequent processings may be possible. Next command processing is performed like the 1st operation gestalt.
[0089] The disk array subset 10 will carry out issue of issue of a data transfer preparation-completion frame, data transfer, and the status frame after normal termination, if preparation of data is completed. While the frame header 401 and the frame payload 402 are changed if needed, a transfer of various frames is performed from the disk array subset 10 before IC2023. FPC20233 of IC2023 receives a data transfer preparation-completion frame, it receives data continuously, stores them in BUF20232, and if a transfer finishes normally continuously, it will report data transfer completion, it receiving a status frame and applying interruption to PTP20231. If interruption is received, PTP20231 starts SPC20230 and directs that it starts data transfer to a host 30. If SPC20230 transmits data to a host 30 and normal termination is checked, it will report normal termination by interruption to PTP20231.
[0090] Here, although Parallel SCSI was shown as an example of host I/F other than a fiber channel, it is possible to apply similarly to ESCON which are other interfaces, for example, host I/F to a mainframe. It is possible to make both so-called open systems, such as a mainframe, and a personal computer, a workstation, intermingled in one set of the disk array system 1, and to connect with it by preparing for example, a fiber channel and the host I/F node corresponding to Parallel SCSI and ESCON as a host I/F node 203 of the disk array switch 20. Although the fiber channel is used like the 1st to 3rd operation gestalt as disk array I/F with this operation gestalt, it is possible to use I/F of arbitration also to disk array I/F.
[0091] The approach of the configuration management of the [5th operation gestalt], next the disk array system 1 is explained as the 5th operation gestalt. Drawing 18 is the system configuration Fig. of this operation gestalt. With this operation gestalt, the host 30 is formed four sets. I/F30 between the disk array systems 1 is connected with a host "#0" and "#1" by Parallel SCSI (Ultra2 SCSI) between the fiber channel, the host "#2", and the disk array system 1 between Parallel SCSI (Ultra SCSI), and a host "#3" and the disk array system 1.
[0092] Connection with the disk array switch 20 of Parallel SCSI is made like the 4th operation gestalt. The disk array system 1 has four sets of the disk array subsets 30. Two independent LUs are constituted by four independent LUs and the disk array subset "#1" at the disk array subset "#0", respectively. One integration LU consists of a disk array subset "#2" and "#3." With this operation gestalt, like the 1st operation gestalt, the disk array subset 10 shall be concealed to a host 30, and the frame of a fiber channel shall be changed. LUN assigned to each LU — the order from LU of a disk array subset "#0" — LUN=0, 1 and 2, and ... they are 7 **s to 6.
[0093] Drawing 19 is an example of the screen displayed on the display screen of an administration terminal 5. Drawing is the logical connection configuration screen in which correspondence with host I/F31 and each Logical unit (LU) was shown.
[0094] The relation between the information 3100 about each host I/F31, the information 11000 about each LU110, the disk array subset 10, and LU110 etc. is displayed on the logical connection configuration screen 50. An I/F class, an I/F rate, the status, etc. are contained as information about host I/F31. As information about LU110, a storing subset number, LUN, capacity, RAID level, the status, information, etc. are displayed. By referring to this screen, a manager can manage the configuration of the disk array system 1 easily.
[0095] On the logical connection configuration screen 50, the line currently drawn between host I/F and LU shows accessible LU110 via each host I/F31. It cannot access from the host 30 who connects with the host I/F from host I/F to LU110 by which a line is not drawn. Since the data format to treat differs and a user also changes with hosts 30, it is indispensable on security maintenance to prepare a suitable access restriction. Then, the manager who sets up a system carries out access restriction by whether the access permission between each LU110 and host I/F is given using this screen. In drawing, although LU "#0" is

accessible from host I/F "#0" and "#1", it cannot be accessed from host I/F "#2" and "#3." LU "#4" is accessible only from host I/F "#2."

[0096] In order to realize such an access restriction, access-restriction information is transmitted from the disk array system-configuration-control means 70 to the disk array switch 20. The access-restriction information sent to the disk array switch 20 is distributed to each host I/F node 203, and is registered into DCT2027 of each host I/F node 203. When the inspection command of LU existence existence to LU to which access was restricted is published by the host, it is each host I/F node's 203 inspecting DCT2027, and not answering to an inspection command, or returning an error, and the LU is no longer recognized from a host by him. In the case of a SCSI protocol, generally as an inspection command of LU existence existence, the Test Unit Ready command and the Inquiry command are used. Since read/write is not carried out without this inspection, it is possible to apply a limit of access easily.

[0097] Although access restriction is applied every host I/F31 with this operation gestalt, it is also easily realizable by extending this to apply access restriction every host 30. Moreover, host I/F31, a host 30, or an address space can be pinpointed, and the access restriction according to the classification of a command to which only the lead was told as good and which good, read/write, and good and read/write told that only a light was improper can also be applied. In this case, a host I/F number, Host ID, an address space, a limit command, etc. are specified as access-restriction information, and a limit is set as the disk array switch 20.

[0098] Next, the addition of the new disk array subset 10 is explained. When adding the disk array subset 10 newly, a manager connects the disk array subset 10 added to the disk array I/F node 202 as for which the disk array switch 20 is vacant. Continuously, a manager operates an administration terminal 5 and does the depression of the carbon button 5001 "reflecting the newest condition" currently displayed on the logical connection configuration screen 50. This actuation is answered and the picture showing a non-set up disk array subset is displayed on a screen (not shown). If it carries out by choosing the picture of this disk array subset, the setting screen of a disk array subset will appear. A manager carries out various setup of the disk array subset added newly on the displayed setting screen. There are a configuration of LU, RAID level, etc. in the item set up here. Continuously, if it changes to the screen of the logical connection block diagram of drawing 19 , a new disk array subset and LU will appear. Henceforth, if the access restriction which receives every host I/F31 is set up and the depression of the "setting activation" carbon button 5002 is carried out, to the disk array switch 20, the information on access-restriction information and a disk array subset, and LU will be transmitted, and a setup will be performed.

[0099] It is performed by the procedure which also mentioned above the procedure at the time of adding LU110 to each disk array subset 10. Moreover, it is performed by a disk array subset and the procedure with the same almost said of deletion of LU. A different point is a point performed, after a manager chooses each deletion part on a screen, and pushes "deletion" carbon button 5003 and a suitable check is performed. As mentioned above, a manager can manage the whole disk array system unitary by using an administration terminal 70.

[0100] Processing of the [6th operation gestalt], next mirroring by the disk array switch 20 is explained as the 6th operation gestalt. Mirroring explained here is the approach of supporting duplex writing by two independent LUs of two sets of disk array subsets, and is the doubleness which even the controller of a disk array subset included. Therefore, dependability differs from doubleness of only a disk.

[0101] The structure of a system in this operation gestalt is the same as what is shown in drawing 1 . the configuration shown in drawing 1 — it shall be, a disk array subset "#0" and "#1" shall completely be equipped with the same LU configuration, and these two disk array subsets shall be seen as one disk array from a host 30 For convenience, the number of the pair of the disk array subset by which mirroring was carried out is called "#01." Moreover, a mirroring pair is formed of LU "#0" and LU "#1" of each disk array subset, and the pair of this LU is called LU "#01" for convenience by them. "Mirrored" is set as CLU Class and, as for the information for managing LU#01 on the host LU configuration table 20271 of DCT2027, the information about LU#0 and LU#1 is set up as LU Info. The configuration of other each part is the same as that of the 1st operation gestalt.

[0102] Actuation of each part in this operation gestalt is the same as that of the 1st example almost. Hereafter, the point which is different from the 1st operation gestalt is explained focusing on actuation of the host I/F node 203 of the disk array switch 20. The mimetic diagram showing the sequence of the frame to which drawing 20 is transmitted at the time of the light actuation in this operation gestalt, drawing 21 , and 22 are flow charts which show the flow of processing by the host I/F node 203 at the time of light actuation.

[0103] The light command frame (FCP_CMD) which the host 30 published is received by IC2023 at the time of light actuation (the arrow head of drawing 20 (a): step 21001). The light command frame received by IC2023 is step 20002 at the time of the lead actuation explained with the 1st operation gestalt. It is

processed like 20005 (step 21002-21005).
[0104] SC2022 searches DCT2027 using SP2021, and recognizes that it is a light access request to LU "#01" of the mirror-ized disk array subset "#01" (step 21006). SC2022 creates the duplicate of the command frame which received on FB2025 (step 21007). SC2022 changes a command frame based on the configuration information set as DCT2027, and creates the separate command frame of both LU "#0" and LU "#1" (step 21008). Here, main LU and LU"#1" is called a main command frame and ** command frame for LU "#0" also to ** LU, a call, and a command frame, respectively. and both — exchange information is separately stored in ET2026, and the command frame created to the disk array subset "#0" and the disk array subset "#1" is published (the arrow head of drawing 20 (b0) (b1): step 21009).
[0105] Each disk array subset "#0" and "#1" receive a command frame, and they transmit a data transfer preparation-completion frame (FCP_XFER_RDY) to the disk array switch 20 independently, respectively (arrow head of drawing 20 (c0) (c1)). Step 20011 of lead actuation [in / by the disk array switch 20 / in the host I/F node 203 / the 1st operation gestalt] The data transfer preparation-completion frame transmitted by the same processing as 20013 is processed (step 21011-21013).
[0106] In the phase which had complete set of data transfer preparation-completion frame from each disk array subset, (step 21014) and SC2022 carry out conversion to the main data transfer preparation-completion frame (step 21015), and transmit the frame after conversion to a host 30 by IC2023 (the arrow head of drawing 20 (d): step 21015).
[0107] A host 30 transmits a data frame (FCP_DATA) to the disk array switch 20 for light data transmission, after receiving a data transfer preparation-completion frame (arrow head of drawing 20 (e)). If the data frame from a host 30 is received by IC2023 (step 21031), like a lead command frame or a light command frame, it will be stored in FB2025 and CRC inspection and analysis of a frame header will be performed (steps 21032 and 21033). Based on the analysis result of a frame header, ET2026 is searched by SP2021 and exchange information is acquired (step 21034).
[0108] SC2022 creates a duplicate like the time of a light command frame (step 21035), one of these is turned to LU "#0" in a disk array subset "#0", and turns another side to LU "#1" in a disk array subset "#1", and transmits (the arrow head of drawing 20 (f0) (f1): step 21037).
[0109] A disk array subset "#0" and "#1" receive a data frame, it carries out a light to a disk unit 104, respectively, and they transmit a status frame (FCP_RSP) to the disk array switch 20.
[0110] SC2022 — a disk array subset "#0" and "#1" — respectively — since — if a status frame is received, an extended header will be removed from those status frames, a frame header will be reproduced, and exchange information will be acquired from ET2026 (steps 21041 and 21042).
[0111] If the status frame from both disk array subset "#0" and "#1" gathers (step 21043), conversion to the main status frame from LU "#0" will be performed after a check of that the status is normal termination (step 21044), and ** status frame elimination will be carried out (step 21045). And IC2023 transmits the command frame for reporting normal termination to a host (the arrow head of drawing 20 (h): step 21046). Finally SP2021 eliminates the exchange information on ET2026 (step 21047).
[0112] The light processing in a mirroring configuration is completed above. Although lead processing to LU "#01" by which mirroring was carried out is performed almost like the light processing which the directions of a data transfer only differ and was mentioned above, unlike a light, it does not need to publish a lead command to two sets of disk array subsets, and should just publish a command frame to either. For example, because of improvement in the speed, although a command frame may always be published to the main LU, if a load is distributed by publishing a command frame by turns etc. to LU of the Lord/******, it is effective.
[0113] In the processing mentioned above, a synchronization of waiting and both is taken for the response of two sets "#0" of disk array subsets, and "#1" at step 21014 and step 21043, and processing is advanced. In such control, after a success of processing by both disk array subsets is checked, in order that processing may progress, the correspondence at the time of error generating becomes easy. On the other hand, in order that the whole processing speed may be dependent on the response of which or the later one, there is a fault that the engine performance falls.
[0114] Since this problem is solved, when it progresses to the next processing, without waiting for the response of a disk array subset or there is a response from one of the disk array subsets, in a disk array switch, it is also possible to carry out "asynchronous" control which progresses to the next processing. In drawing 20 , a broken-line arrow head shows an example of the frame sequence at the time of performing asynchronous control.
[0115] In the frame sequence shown by the broken-line arrow head, transmission of the data transfer preparation-completion frame to the host to whom it is carried out at step 21016 is carried out after processing of step 21009, without waiting for the data transfer preparation-completion frame from the disk array subset 10. In this case, the data transfer preparation-completion frame transmitted to a host is

generated by SC2022 of the disk array switch 20 (broken-line arrow head (d')).

[0116] From a host 30, a data frame is transmitted to the disk array switch 20 to the timing shown by the broken-line arrow head (e'). With the disk array switch 20, this data frame is once stored in FB2025. SC2022 answers reception of the data transfer preparation-completion frame from the disk array subset 10, and transmits the data frame held at FB2025 to the disk array subset 10 to which the data transfer preparation-completion frame has been sent (a broken-line arrow head (f0'), (f1')).

[0117] The termination report to a host 30 from the disk array switch 20 is performed when there is a report (a broken-line arrow head (g0'), (g0')) from both disk array subsystems 10 (broken-line arrow head (h')). It is possible only for the part of the time amount Ta shown in drawing 20 to shorten the processing time by such processing.

[0118] The following processings are carried out when an error occurs in the middle of the frame transfer between the disk array switch 20 and the disk array subset 10.

[0119] When the processing under activation is light processing, retry processing is performed to LU which the error generated. If a retry is successful, processing will be continued as it is. When the retry of the count of the convention set up beforehand goes wrong, the disk array switch 20 forbids access to this disk array subset 10 (or LU), and registers into DCT2027 the information which shows that. Moreover, the disk array switch 20 notifies that to the disc system configuration means 70 via MP200 and the communication link controller 204.

[0120] The disc system configuration means 70 answers this notice, and publishes an alarm to an administration terminal 5. Thereby, a manager can recognize that the trouble occurred. Then, the disk array switch 20 continues operation using a normal disk array subset. A host 30 does not recognize that the error occurred and can continue processing.

[0121] raising the failure-proof nature of a disk, since a mirror configuration is realizable with two sets of disk array subsystems according to this operation gestalt — things are made. Moreover, a disk array controller, disk array I/F, and the failure-proof nature of a disk array I/F node can be raised, there is nothing with ** which carries out doubleness of an internal bus etc., and the dependability of the whole disk array system can be raised.

[0122] Three or more sets of the [7th operation gestalt], next the disk array subsets 10 are unified, and how to constitute the group of one set of a logical disk array subset is explained. Data are distributed and stored in two or more disk array subsets 10 with this operation gestalt. Thereby, access to a disk array subset is distributed and a total throughput is raised by inhibiting concentration of access to a specific disk array subset. With this operation gestalt, such striping processing is carried out with a disk array switch.

[0123] Drawing 23 is the address map of the disk array system 1 in this operation gestalt. Striping of the address space of the disk array subset 10 is carried out in the SUTOREIPU size S. The address space of the disk array system 1 seen from the host is distributed by every stripe size S a disk array subset "#0", "#1", "#2", and "#3." Although the size of the stripe size S is arbitrary, the direction which is not not much small is good. When the stripe size S was too small and the stripe crossover to which the data which should be accessed belong to two or more stripes occurs, a possibility that an overhead may occur is in the processing. If stripe size S is enlarged, since the probability for a stripe crossover to occur will decrease, for the improvement in the engine performance, it is desirable. The number of LUs can be set as arbitration.

[0124] Hereafter, actuation of the host I/F node 203 in this operation gestalt is explained and explained paying attention to difference with the 1st operation gestalt, referring to the operation flow chart shown in drawing 24 . In addition, with this operation gestalt, "Striped" is set to CLU Class of the information about the host LU by whom striping was done on the host LU configuration table 20271 of DCT2027, and stripe size "S" is set to CLU Stripe Size.

[0125] If a host 30 publishes a command frame, the disk array switch 20 will recognize that SC2022 which receives this (step 22001) needs to search DCT2027, and needs to carry out striping of this command frame using reception and SP2021 from IC2023 by IC2023 of the host I/F node 203 (step 22005).

[0126] Next, SC2022 searches DCT2027 by SP2021, asks for the stripe number of the stripe with which the data set as the object of access belong from the configuration information containing the stripe size S, and specifies in which disk array subset 10 this stripe is stored (step 22006). Under the present circumstances, although a stripe crossover may occur, about the processing in this case, it mentions later. When a stripe crossover does not occur, based on SP's2021 count result, SC2022 changes to a command frame (step 22007), and stores exchange information in ET2026 (step 22008). Henceforth, the same processing as the 1st operation gestalt is performed.

[0127] When a stripe crossover occurs, SP2021 generates two command frames. This generation is performed with reproducing the command frame which the host 30 published. The frame header of the command frame to generate, a frame payload, etc. are set up newly. Although it is also possible like the 6th operation gestalt to carry out conversion after creating the duplicate of a command frame by SC2022, it shall be newly created by SP2021 here. SC2022 will transmit these to each disk array subset 10, if two command frames are generated.

[0128] Then, data transfer is carried out like the 1st operation gestalt. Here, unlike the 1st operation gestalt or the 6th operation gestalt, it is necessary to transmit the data itself with this operation gestalt in one set of the two sets of a host 30 and the disk array subsets 10. For example, in lead processing, it is necessary to transmit all the data frames transmitted from two sets of the disk array subsets 10 to a host 30. Under the present circumstances, to the data frame transmitted from each disk array subset 10, according to the exchange information registered into ET2026, SC2022 is suitable sequence, adds suitable exchange information and transmits to a host 30.

[0129] In light processing, it transmits to the disk array subset 10 which divides and corresponds to two data frames like the case of a command frame. In addition, the sequence control of a data frame is not indispensable if the host or the disk array subset supports the processing in random order called an AUTOOBU order (Out of Order) function.

[0130] If all data transfer is completed and the disk array switch 20 finally receives two status frames from the disk array subset 10, SP2021 (or SC2022) will create the status frame to a host 30, and will transmit this to a host 30 by IC2023.

[0131] Since access can be distributed to two or more disk array subsets, while being able to raise a throughput as total according to this operation gestalt, also as for an access latency, it is possible to make it decrease on the average.

[0132] Creation of the duplicate for two sets (or disk array subset) of the [8th operation gestalt], next disk array systems is explained as the 8th operation gestalt. A system which is explained here arranges one side of two sets of disk array systems to a remote place, and is equipped with the resistance over the failure of the disk array system of another side by a natural disaster etc. The thing of creation of the duplicate performed between a disaster recovery, and a call and the disk array system of a remote place in a cure to such a disaster is called a remote copy.

[0133] Since mirroring explained with the 6th operation gestalt constitutes a mirror from the disk array subset 10 installed in the geographical almost same location, disk array I/F21 is good by the fiber channel. However, when the disk array (disk array subset) which performs a remote copy is installed in the remote place exceeding 10km, you cannot transmit a frame without junction by the fiber channel. Since [this] the distance between each other is usually set to hundreds of km or more when used for a disaster recovery, it is impossible practically to connect between disk arrays by the fiber channel, and a high-speed public line, satellite communication, etc. by ATM (Asynchronous Transfer Mode) etc. are used.

[0134] Drawing 25 is an example of the disaster recovery structure of a system in this operation gestalt.

[0135] 81 is Site A, 82 is Site B, and both sites are installed in a geographical remote place. 9 is a public line and an ATM packet passes through this. A site A81 and a site B82 have the disk array system 1, respectively. Here a site A81 is a common site usually used, and a site B82 is a remote disaster recovery site used when a site A81 is downed with disaster etc.

[0136] The contents of the disk array subset "#0" of the disk array system 10 of a site A81 and "#1" are copied to the disk array subset for a remote copy of the disk array system 10 of a site B82 "#0", and "#1." What is connected to a remote site among the I/F nodes of the disk array switch 20 is connected to the public line 9 using ATM. This node is called the ATM node 205. The ATM node 205 is constituted like the host I/F node shown in drawing 5 , and IC2023 changes an ATM-fiber channel. This conversion is realized by the same approach as conversion of the SCSI-fiber channel in the 4th operation gestalt.

[0137] Processing of the remote copy in this operation gestalt is similar with processing of mirroring in the 6th operation gestalt. Hereafter, a different point from processing of mirroring in the 6th operation gestalt is explained.

[0138] If a host 30 publishes a light command frame, the disk array system 10 of a site A81 will double a frame like the case in the 6th operation gestalt, and will transmit one of these to the own disk array subset 10. The frame of another side is changed into an ATM packet from a fiber channel frame by the ATM node 205, and is sent to a site B82 through a public line 9.

[0139] To a site B82, the ATM node 205 of the disk array switch 20 receives this packet. IC2023 of the ATM node 205 reproduces a fiber channel frame from an ATM packet, and transmits it to SC2022. SC2022 performs frame conversion like the time of receiving a light command from a host 30, and transmits it to the disk array subset for a remote copy. Henceforth, in data transfer preparation-completion frames, data frames, and all the status frames, a remote copy is realizable by performing fiber channel-ATM conversion in the ATM node 205, and carrying out same frame transfer processing.

[0140] When a host 30 publishes a lead command frame, the disk array switch 20 transmits a command frame only to the disk array subset 10 of a self-site, and leads data only from the disk array subset 10 of a

self-site. The actuation at this time becomes the same as that of the 1st operation gestalt.
[0141] According to this operation gestalt, user data can be backed up on real time and it can have the resistance over the site failure by a natural disaster etc., and disk array system failure.
[0142] Integration of two or more LUs included by one set of the [9th operation gestalt], next the disk array subset 10 is explained. For example, in order that the disk unit for main frames may maintain compatibility with the past system, the maximum of the size of a logical volume is set as 2GB. When sharing such a disk array system also with an open system, LU will receive a limit of logical volume size as it is, and many its LUs of small size can be seen from a host. By such approach, when large capacity-ization progresses, the problem that employment becomes difficult arises. Then, it considers unifying this logical volume (namely, LU) and constituting one big integration LU by the function of the disk array switch 20. In this operation gestalt, integration LU is created with the disk array switch 20.
[0143] Integration of LU in this operation gestalt is the same as that of creation of the integration LU by two or more disk array subsets 10 which can be set in the 1st operation gestalt. Difference is only integration by the plurality LU in the same disk array subset 10. The actuation as a disk array system becomes completely the same as that of the 1st operation gestalt.
[0144] Thus, by unifying two or more LUs included by the same disk array subset 10, and creating one big LU, it becomes unnecessary to manage many LUs from a host, and excels in operability, and the disk array system which reduced management cost can be built.
[0145] The setting approach of the [10th operation gestalt], next the shift pass by the disk array switch 10 is explained referring to drawing 26 .
[0146] The configuration of each part in the computing system shown in drawing 26 is the same as that of the 1st operation gestalt. Here, if two sets of hosts 30 access the disk array subset 10 using respectively different disk array I/F21, it will be assumed that it constitutes like. By a diagram, only the number which needs the host I/F node 203 and the disk array I/F node 202 of a disk array subset and the disk array switch 20 for explanation here is shown.
[0147] The disk array subset 10 had the same configuration as drawing 2 , and has connected two disk array I/F controllers to one set of the disk array switch 20, respectively. The shift pass of disk array I/F21 is set to DCT227 of each node of the disk array switch 20. Shift pass is pass of the alternative established so that it may become accessible, also when a failure occurs on one certain pass. Here, the shift pass of disk array I/F "#1" and disk array I/F "#1" is set for the shift pass of disk array I/F "#0" as disk array I/F "#0." Similarly, shift pass is set also about each between the high order adapters in the disk array subset 10, between a cache and alternate memory, and between low order adapters.
[0148] Next, as shown in drawing 26 , disk array I/F21 linked to the high order adapter "#1" of the disk array subset 1 is disconnected, it assumes that the failure occurred, and the setting-operation of shift pass is explained. It becomes impossible for the host "#1" using disk array I/F21 which the failure generated to access the disk array subset 10 at this time. It is recognized as the failure having generated the disk array switch 20 on this pass, when not recovering, even if it detected the abnormalities of the frame transfer between the disk array subsets 10 and carried out retry processing.
[0149] If the failure of pass occurs, SP2021 will register that the failure occurred in disk array I/F "#1" into DCT2027, and will register using disk array I/F "#0" as shift pass. Henceforth, SC2022 of the host I/F node 203 operates so that the frame from a host "#1" may be transmitted to the disk array I/F node 202 linked to disk array I/F "#0."
[0150] The high order adapter 101 of the disk array subset 10 succeeds and processes the command from a host "#1." Moreover, the disk array switch 20 notifies generating of a failure to the disk array system-configuration-control means 70, and generating of a failure is notified to a manager by the disk array system-configuration-control means 70.
[0151] According to this operation gestalt, a change on the shift pass at the time of a failure occurring on pass can be performed without making it recognize to a host side, and a shift processing setup by the side of a host can be made unnecessary. Thereby, the availability of a system can be raised.
[0152] Each operation gestalt explained above explained the disk array system which used the disk unit altogether as storage media. However, this invention is not limited to this, and when not only a disk unit but an optical disk unit, a tape unit, DVD equipment, a semiconductor memory, etc. are used as storage media, it can be applied similarly.

[Translation done.]

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1] It is the block diagram of the computer system of the 1st operation gestalt.
[Drawing 2] It is the block diagram of the disk array subset of the 1st operation gestalt.
[Drawing 3] It is the block diagram of the disk array switch of the 1st operation gestalt.
[Drawing 4] It is the block diagram of the crossbar switch of the disk array switch in the 1st operation gestalt.
[Drawing 5] It is the block diagram of the host I/F node of the disk array switch in the 1st operation gestalt.
[Drawing 6] It is the block diagram of a system configuration table.
[Drawing 7] It is the block diagram of a subset configuration table.
[Drawing 8] It is the block diagram of the frame of a fiber channel.
[Drawing 9] It is the block diagram of the frame header of a fiber channel.
[Drawing 10] It is the block diagram of the frame payload of a fiber channel.
[Drawing 11] It is the mimetic diagram showing the sequence of the frame transmitted through a fiber channel at the time of the lead actuation from a host.
[Drawing 12] It is the mimetic diagram showing LU of Host LU and each disk array subset, and the correspondence relation of each disk unit.
[Drawing 13] It is the flow chart of the processing in the host I/F node at the time of light processing.
[Drawing 14] It is the block diagram of a switching packet.
[Drawing 15] It is the disk array structure-of-a-system Fig. which made cluster connection of two or more disk array switches.
[Drawing 16] It is the block diagram of the computer system in the 2nd operation gestalt.
[Drawing 17] It is the block diagram of the interface controller of the disk array switch in the 4th operation gestalt.
[Drawing 18] It is the block diagram of the computer system in the 5th operation gestalt.
[Drawing 19] It is the screen block diagram showing the example of a display of a logical connection configuration screen.
[Drawing 20] It is the mimetic diagram showing the frame sequence in the 6th operation gestalt.
[Drawing 21] It is the flow chart of the processing in the host I/F node at the time of mirroring light processing of the 6th operation gestalt.
[Drawing 22] It is the flow chart of the processing in the host I/F node at the time of mirroring light processing of the 6th operation gestalt.
[Drawing 23] It is the mimetic diagram showing the correspondence relation of Host LU and LU of each disk array subset in the 7th operation gestalt.
[Drawing 24] It is the flow chart which shows processing of the host I/F node in the 7th operation gestalt.
[Drawing 25] It is a disaster recovery structure-of-a-system Fig. in the 8th operation gestalt.
[Drawing 26] It is an explanatory view about a setup of shift pass.
[Description of Notations]
1 [— A disk array switch, 30 / — A host computer, 70 / — A disk array system-configuration-control means, 200 / — A management processor, 201 / — A crossbar switch, 202 / — A disk array I/F node, 203 / — A host I/F node, 204 / — Communication link controller.] — A disk array system, 5 — An administration terminal, 10 — A disk array subset, 20

[Translation done.]

* NOTICES *

JPO and NCIPI are not responsible for any damages caused by the use of this translation.

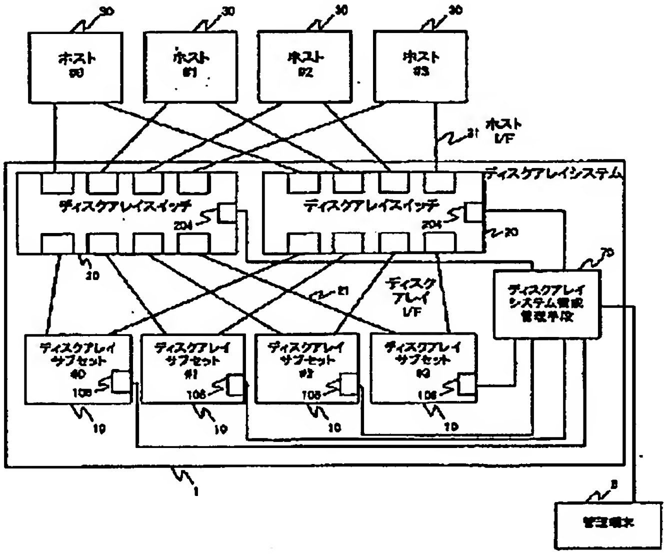
1.This document has been translated by computer. So the translation may not reflect the original precisely.

2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DRAWINGS

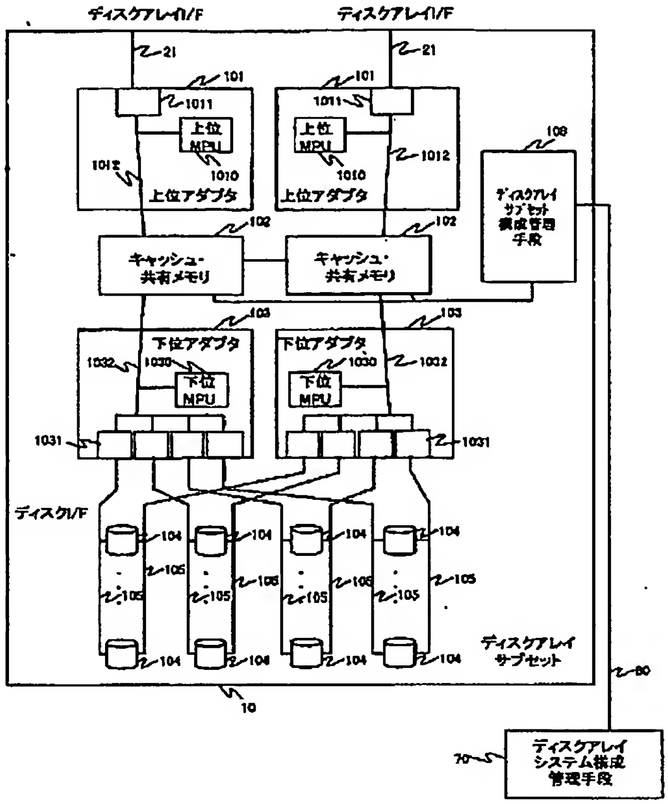
[Drawing 1]

図1


30
ホスト #0
ホスト #1
ホスト #2
ホスト #3
31
ホスト I/F
ディスクアレイシステム
ディスクアレイスイッチ
204
20
21
ディスクアレイ I/F
70
ディスクアレイシステム構成管理手段
ディスクアレイサブセット #0
ディスクアレイサブセット #1
ディスクアレイサブセット #2
ディスクアレイサブセット #3
106
10
1
8
管理端末

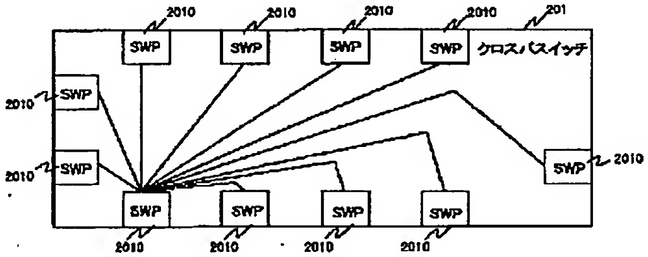
[Drawing 2]

図2



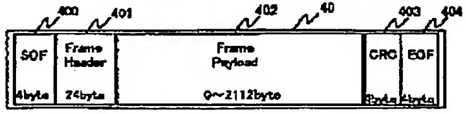
[Drawing 4]

図4



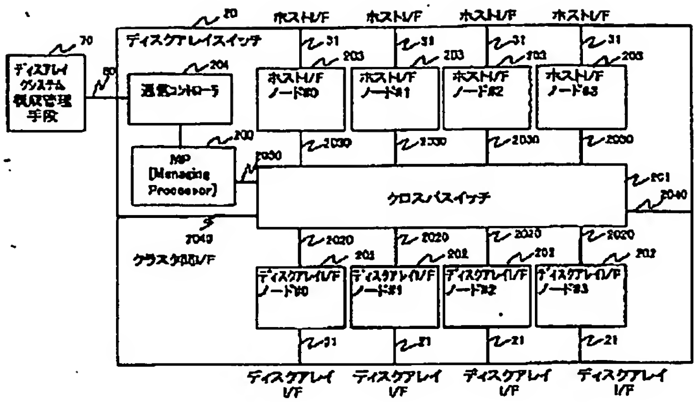
[Drawing 8]

図8



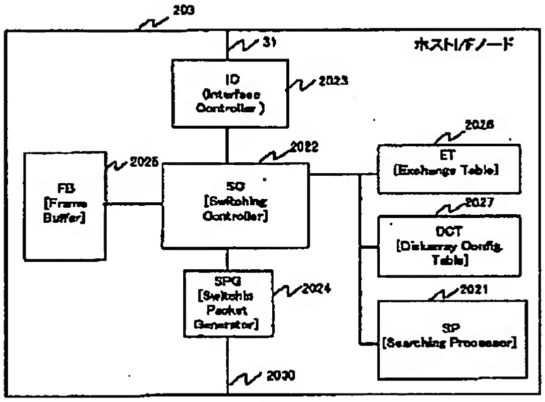
[Drawing 3]

図3



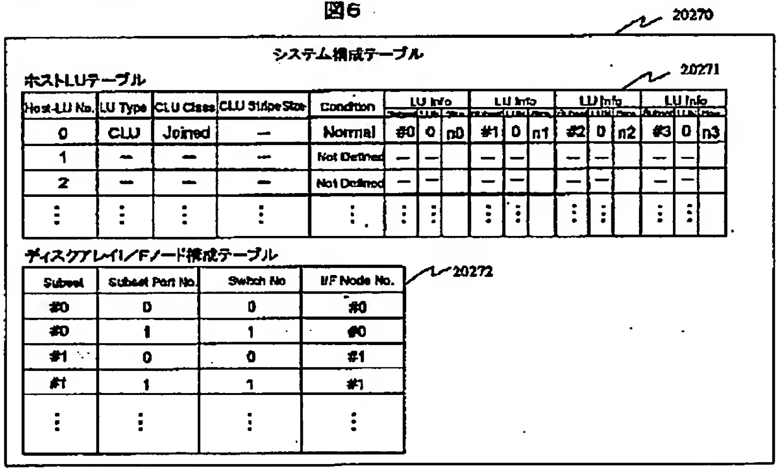
[Drawing 5]

図5



[Drawing 6]

図6


20270
20271
20272

システム構成テーブル

ホストLUテーブル

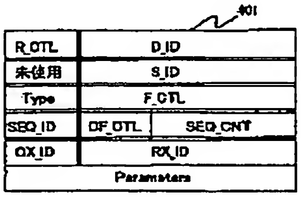
Host-LU No.	LU Type	CLU Class	CLU Stripe Size	Condition	LU Info			LU Info			LU Info			LU Info		
					Subset	LUN	Size	Subset	LUN	Size	Subset	LUN	Size	Subset	LUN	Size
0	CLU	Joined	—	Normal	#0	0	n0	#1	0	n1	#2	0	n2	#3	0	n3
1	—	—	—	Not Defined	—	—		—	—		—	—		—	—	
2	—	—	—	Not Defined	—	—		—	—		—	—		—	—	
⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮		⋮	⋮		⋮	⋮	

ディスクアレイI/Fノード構成テーブル

Subset	Subset Port No.	Switch No	I/F Node No.
#0	0	0	#0
#0	1	1	#0
#1	0	0	#1
#1	1	1	#1
⋮	⋮	⋮	⋮

[Drawing 9]

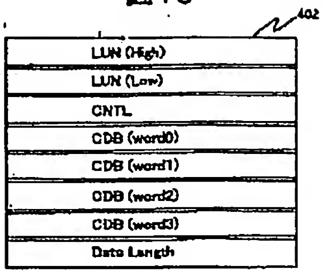
図9



R_CTL	D_ID	
未使用	S_ID	
Type	F_CTL	
SEQ_ID	DF_CTL	SEQ_CNT
OX_ID	RX_ID	
Parameters		

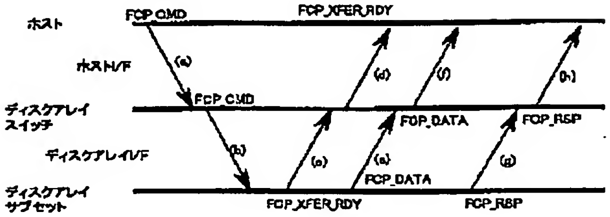
[Drawing 10]

図10



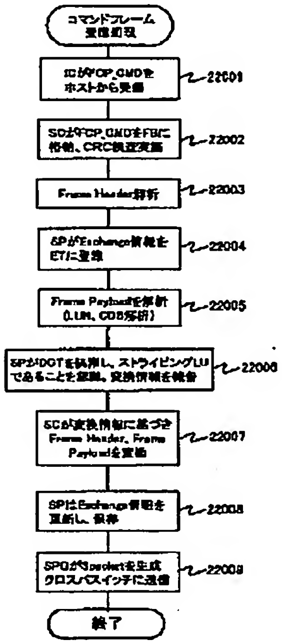
[Drawing 11]

図11



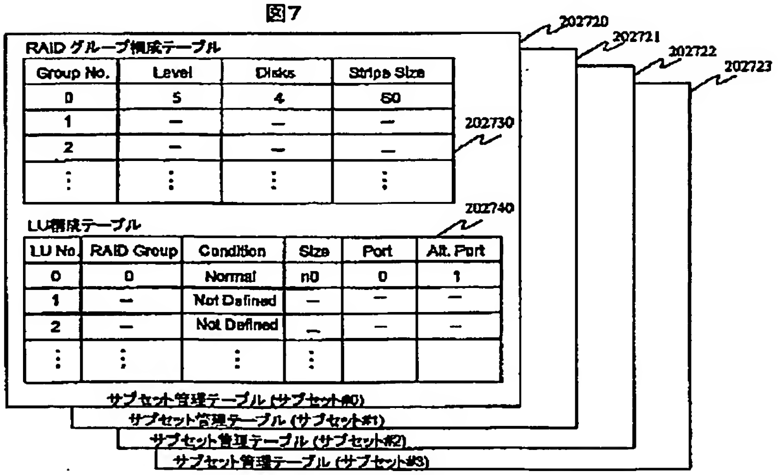
[Drawing 24]

図24


コマンドフレーム受信処理
SCがFCP_CMDをホストから受信
22001
SCはFCP_CMDをFBに格納、CRC検査実施
22002
Frame Header解析
22003
SPがExchange情報をETに登録
22004
Frame Payloadを解析(LUN、CDB解析)
22005
SPがDDTを参照し、ストライピングLUであることを認識、変換情報を検索
22006
SCが変換情報に基づきFrame Header、Frame Payloadを変換
22007
SPはExchange情報を更新し、保持
22008
SPCがSpacketを生成、クロスパスイッチに送信
22009
終了

[Drawing 7]

図7


202720
202721
202722
202723
202730
202740
サブセット管理テーブル (サブセット#0)
サブセット管理テーブル (サブセット#1)
サブセット管理テーブル (サブセット#2)
サブセット管理テーブル (サブセット#3)

RAID グループ構成テーブル

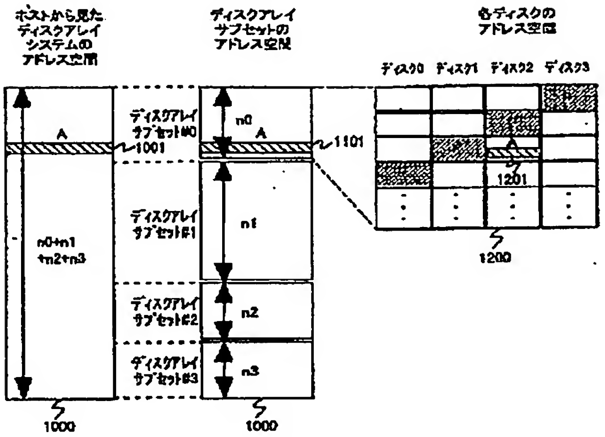
Group No.	Level	Disks	Stripe Size
0	5	4	50
1	—	—	—
2	—	—	—
⋮	⋮	⋮	⋮

LU構成テーブル

LU No.	RAID Group	Condition	Size	Port	Alt. Port
0	0	Normal	n0	0	1
1	—	Not Defined	—	—	—
2	—	Not Defined	—	—	—
⋮	⋮	⋮	⋮		

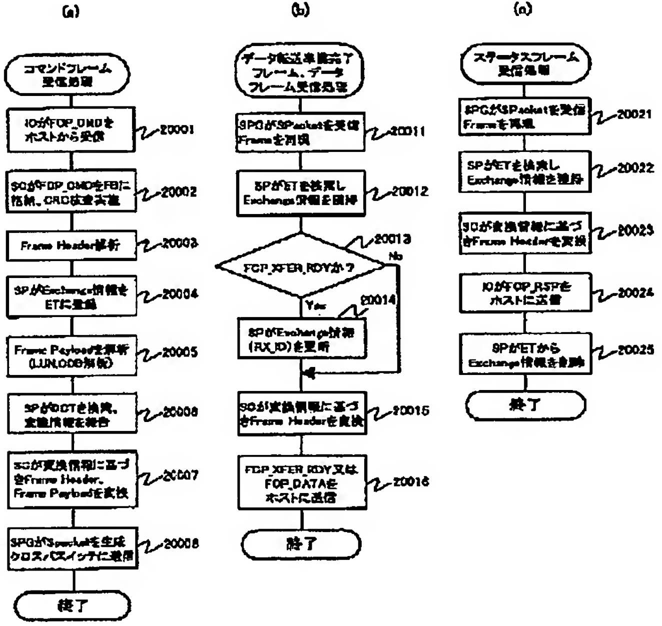
[Drawing 12]

図12


ホストから見たディスクアレイシステムのアドレス空間
ディスクアレイサブセットのアドレス空間
各ディスクのアドレス空間
ディスク0
ディスク1
ディスク2
ディスク3
A
n0+n1+n2+n3
1001
1000
ディスクアレイサブセット#0
ディスクアレイサブセット#1
ディスクアレイサブセット#2
ディスクアレイサブセット#3
n0
A
n1
n2
n3
1101
1000
1201
1200

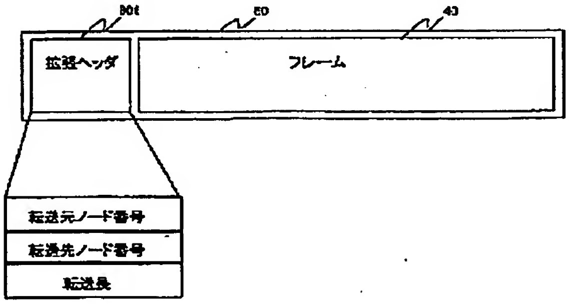
[Drawing 13]

図13


(a)
コマンドフレーム受信処理
IOがFCP_CMDをホストから受信
20001
SCがFCP_CMDをFBに格納、CRC検査実施
20002
Frame Header解析
20003
SPがExchange情報をETに登録
20004
Frame Payloadを解析(LUN,CDB解析)
20005
SPがCTを検索、変換情報を獲得
20006
SCが変換情報に基づきFrame Header、Frame Payloadを変換
20007
SPGがSpacketを生成クロスパスイッチに送信
20008
終了
(b)
データ転送準備完了フレーム、データフレーム受信処理
SPGがSPacketを受信Frameを再現
20011
SPがETを検索しExchange情報を獲得
20012
FCP_XFER_RDYか？
20013
No
Yes
20014
SPがExchange情報(RX_ID)を更新
SCが変換情報に基づきFrame Headerを変換
20015
FCP_XFER_RDY又はFCP_DATAをホストに送信
20016
終了
(c)
ステータスフレーム受信処理
SPGがSPacketを受信Frameを再現
20021
SPがETを検索しExchange情報を獲得
20022
SCが変換情報に基づきFrame Headerを変換
20023
IOがFCP_RSPをホストに送信
20024
SPがETからExchange情報を削除
20025
終了

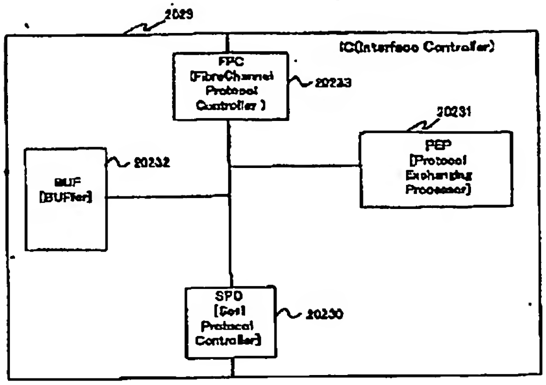
[Drawing 14]

図14



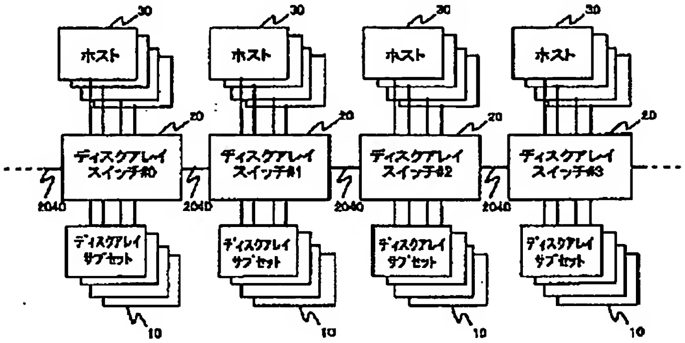
[Drawing 17]

図17



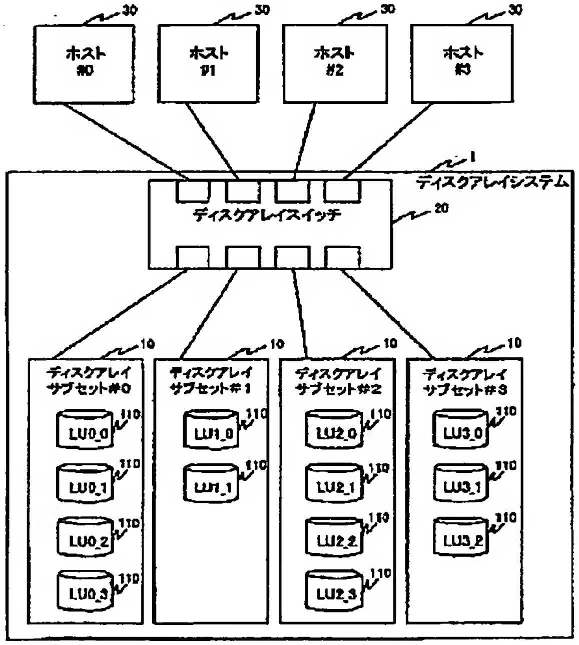
[Drawing 15]

図15



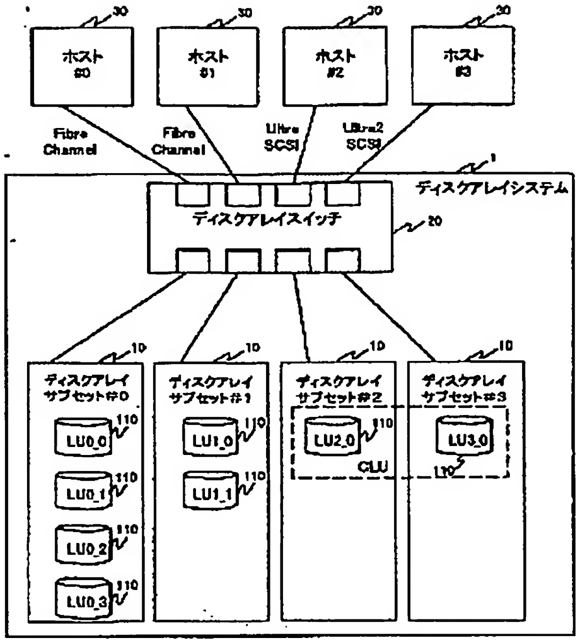
[Drawing 16]

図16



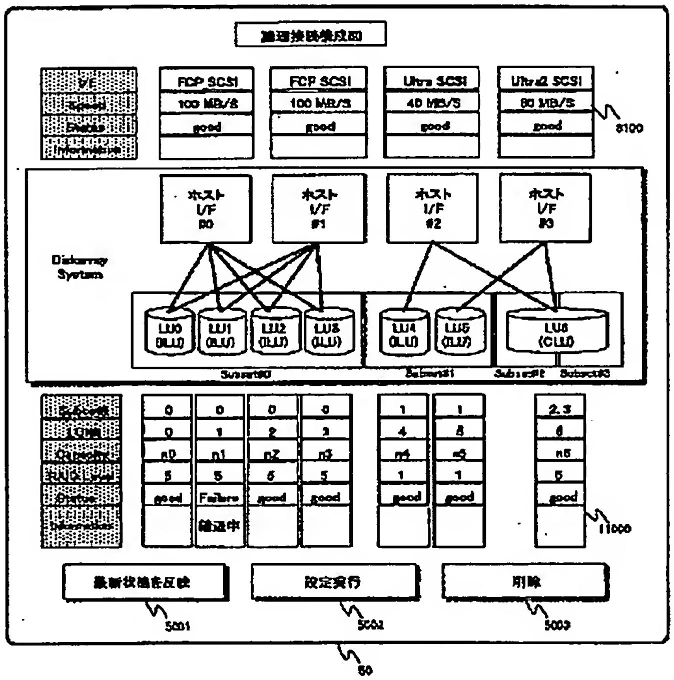
[Drawing 18]

図18



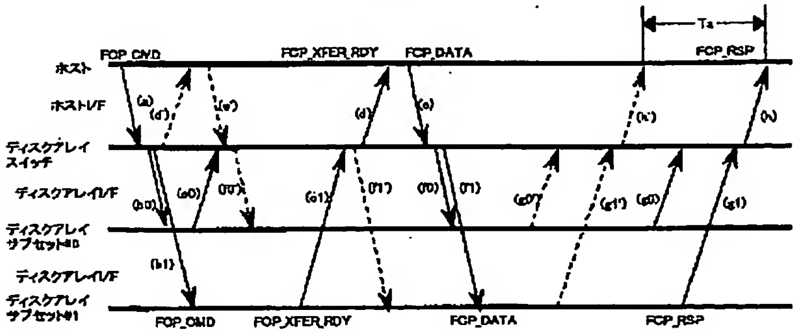
[Drawing 19]

図19



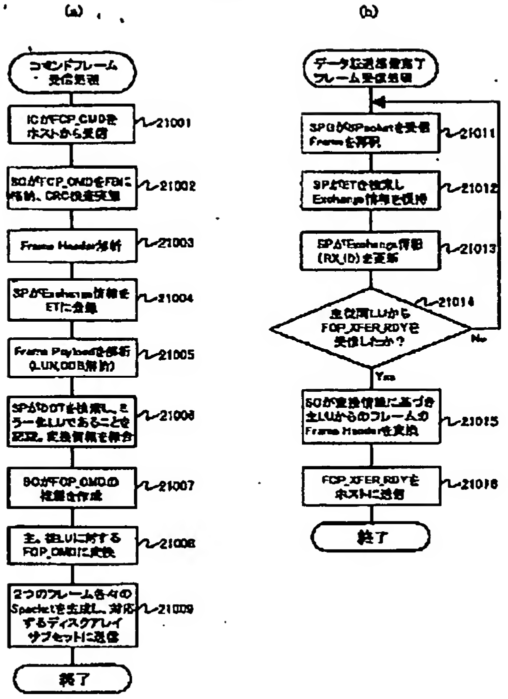
[Drawing 20]

図20



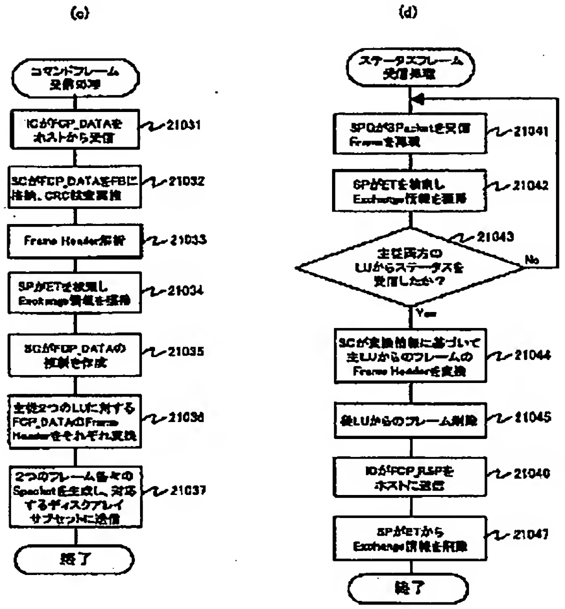
[Drawing 21]

図21


(a)
コマンドフレーム受信処理
ICがFCP_CMDをホストから受信
21001
SCがFCP_CMDをFBに格納、CRC検査実施
21002
Frame Header解析
21003
SPがExchange情報をETに登録
21004
Frame Payloadを解析(LUN,CDB解析)
21005
SPがDDTを検索し、ミラー化LUであることを認識。変換情報を併合
21006
SCがFCP_CMDの複製を作成
21007
主、従LUに対するFCP_CMDに変換
21008
2つのフレーム各々のSpacketを生成し、対応するディスクアレイサブセットに送信
21009
終了
(b)
データ転送準備完了フレーム受信処理
SPGがSPacketを受信Frameを再現
21011
SPがETを検索しExchange情報を獲得
21012
SPがExchange情報(RX_ID)を更新
21013
主従両LUからFCP_XFER_RDYを受信したか？
21014
No
Yes
SCが変換情報に基づき主LUからのフレームのFrame Headerを変換
21015
FCP_XFER_RDYをホストに送信
21016
終了

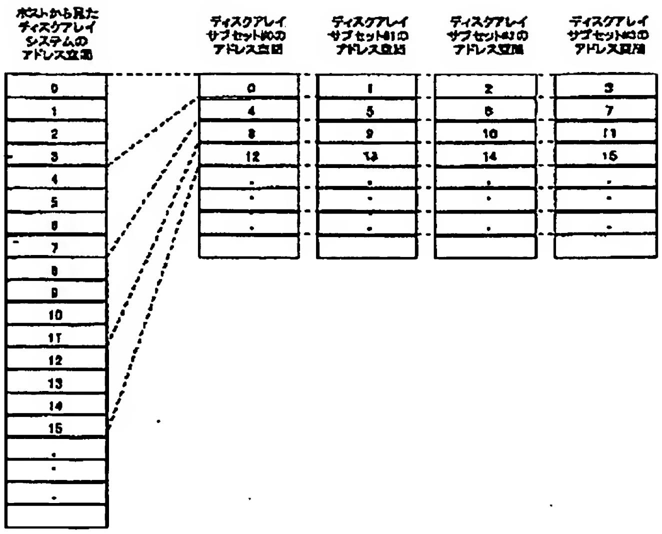
[Drawing 22]

図22


(c)
コマンドフレーム受信処理
ICがFCP_DATAをホストから受信
21031
SCがFCP_DATAをFBに格納、CRC検査実施
21032
Frame Header解析
21033
SPがETを検索しExchange情報を獲得
21034
SCがFCP_DATAの複製を作成
21035
主従2つのLUに対するFCP_DATAのFrame Headerをそれぞれ変換
21036
2つのフレーム各々のSpacketを生成し、対応するディスクアレイサブセットに送信
21037
終了
(d)
ステータスフレーム受信処理
SPGがSPacketを受信Frameを再現
21041
SPがETを検索しExchange情報を獲得
21042
主従両方のLUからステータスを受信したか？
21043
No
Yes
SCが変換情報に基づいて主LUからのフレームのFrame Headerを変換
21044
従LUからのフレーム削除
21045
ICがFCP_RSPをホストに送信
21046
SPがETからExchange情報を削除
21047
終了

[Drawing 23]

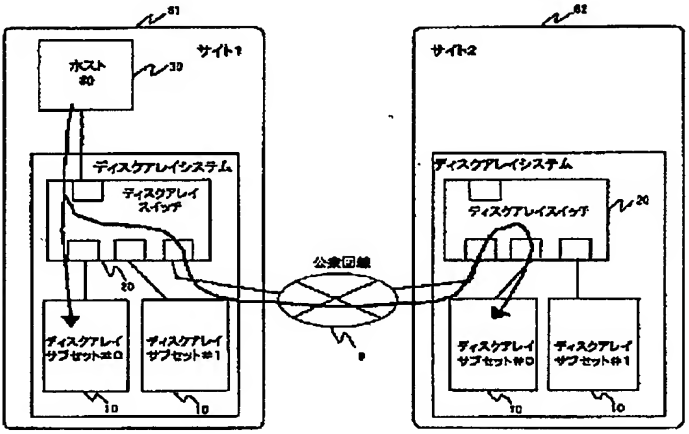
図23



ホストから見たディスクアレイシステムのアドレス空間
ディスクアレイサブセット#0のアドレス空間
ディスクアレイサブセット#1のアドレス空間
ディスクアレイサブセット#2のアドレス空間
ディスクアレイサブセット#3のアドレス空間
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
0
4
8
12
1
5
9
13
2
6
10
14
3
7
11
15

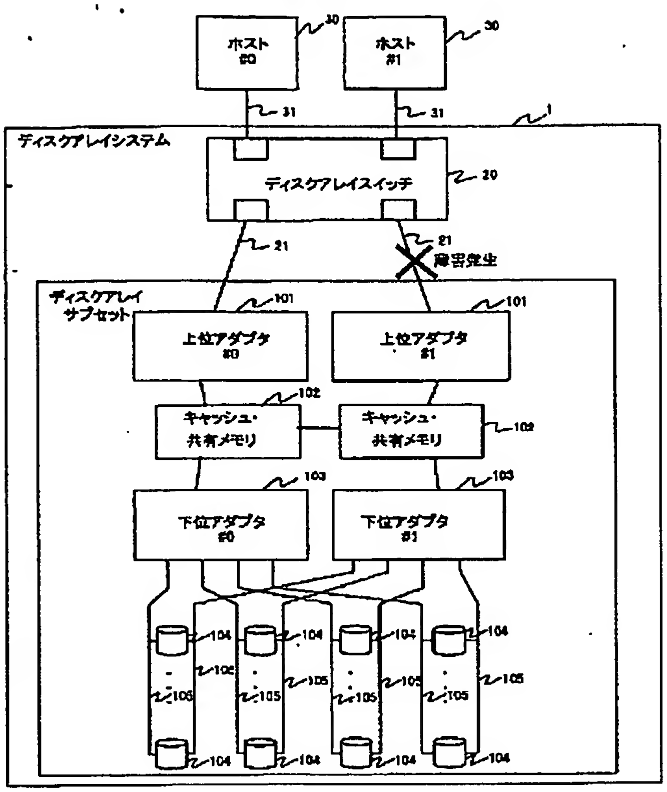
[Drawing 25]

図25



[Drawing 26]

図26



[Translation done.]

CORRECTION OR AMENDMENT

[Kind of official gazette] Printing of amendment by the convention of 2 of Article 17 of Patent Law
[Section partition] The 3rd partition of the 6th section
[Publication date] December 9, Heisei 16 (2004. 12.9)

[Publication No.] JP,2000-242434,A (P2000-242434A)
[Date of Publication] September 8, Heisei 12 (2000. 9.8)
[Application number] Japanese Patent Application No. 11-344260
[The 7th edition of International Patent Classification]

G06F 3/06

[FI]

G06F 3/06 301 G
G06F 3/06 540

[Procedure revision]
[Filing Date] December 24, Heisei 15 (2003. 12.24)
[Procedure amendment 1]
[Document to be Amended] Specification
[Item(s) to be Amended] Claim
[Method of Amendment] Modification
[The contents of amendment]
[Claim(s)]
[Claim 1]
It is the switch connected to two or more storage systems which have a host computer and the controller which controls access to a disk and a disk respectively,
The first node connected with a host computer,
It has two or more second nodes respectively connected with a storage system,
Said first node receives the first access request from a host computer,
Said switch is a switch characterized by changing into the second access request of addressing to the first Logical unit which is a storage region in the disk with which the storage system connected to the second node has the first access request which received, and transmitting this second access request to this storage system through the second node connected with the storage system which has this first Logical unit.
[Claim 2]
It is a switch according to claim 1,
Said first access request has the first address which shows the storing location of the data for access within the second Logical unit which consists of two or more first Logical unit,
For said switch, said two or more first Logical unit is the switches characterized by changing into the second address which shows the storing location of the data for access inside, and transmitting the second access request which has this second address either about said first address.
[Claim 3]
It is a switch according to claim 2,
It is the switch which said first access request has the frame payload, and is characterized by storing said first address in a frame payload.
[Claim 4]
It is a switch according to claim 3,

The switch characterized by containing a logical unit number and a logical block address in said first address.

[Claim 5]

It is a switch according to claim 1,

Said first access request is a switch characterized by being the access request of addressing to the second Logical unit which consists of two or more first Logical unit containing the first Logical unit which stores the data for access.

[Claim 6]

It is a switch according to claim 1,

Said the first access request and said second access request have the frame header in which the frame destination ID is stored,

The first frame destination ID which shows the first node is included in said first access request,

Said switch is a switch characterized by changing into the second frame destination ID which shows either of two or more storage systems by which said first frame destination ID is respectively connected to the second node, and transmitting the second access request which has the second frame destination ID after conversion to either of said two or more storage systems.

[Claim 7]

a switch according to claim 1 — it is — further

The switch characterized by having the configuration information of the storage region which two or more storage systems respectively connected to a switch through the second node have.

[Claim 8]

They are claim 2 or a switch according to claim 5,

Said second Logical unit is a switch characterized by consisting of two or more first Logical unit which exists in a respectively different storage system.

[Claim 9]

It is the switch connected to two or more store systems which have a calculating machine and the control section respectively connected to two or more disks and these two or more disks,

1 or two or more first nodes which are connected with a computer,

It has two or more second nodes respectively connected to a storage system,

The first node receives the first access request from a computer,

Said switch is changed into the second access request addressed to the Logical unit which consists of storage regions in two or more disks where either of said two or more store systems has the first access request,

The second node is a switch characterized by transmitting this second access request to the store system which has this Logical unit.

[Claim 10]

It is the system which stores the data accessed from a calculating machine,

respectively — 1 or two or more first logic storage regions — this — two or more storage subsystems which have the control section which controls access to 1 or two or more first logic storage regions,

It has the switch connected with said two or more storage subsystems and calculating machine,

Said switch,

The first access request is received from a computer,

The first access request which received is changed into the second access request addressed to the first logic storage region,

The system characterized by transmitting the second access request to either of said two or more storage subsystems.

[Claim 11]

It is a system according to claim 10,

Said first logic storage region is a system characterized by being the storage region which consists of storage regions in two or more disks which a storage subsystem has.

[Claim 12]

It is a system according to claim 10,

Said first access request has the first address which shows the second logic storage region which has two or more first logic storage regions,

Said switch is a system characterized by changing said first address into the second address which shows either of said two or more first logic storage regions, and transmitting the second access request which has this second address to either of said two or more storage subsystems.

[Claim 13]

It is a system according to claim 12,

Said second logic storage region is a system characterized by consisting of two or more first logic storage regions which exist in a respectively different storage system.

[Claim 14]

It is a system according to claim 10,

Said first access request has the first address which shows the storing location of the data for access in the second logic storage region which consists of two or more first logic storage regions,

For said switch, said two or more first logic storage regions are the systems characterized by changing into the second address which shows the storing location of the data for access inside, and transmitting the second access request which has this second address to either of said two or more storage subsystems either about said first address.

[Claim 15]

They are claim 12 thru/or a system according to claim 14,

It is the system which said first access request and said second access request have the frame payload, and is characterized by storing respectively said first address and said second address in a frame payload.

[Claim 16]

They are claim 12 thru/or a system according to claim 15,

The system characterized by containing a logical unit number in said first access request.

[Claim 17]

It is a system according to claim 10,

Said switch has two or more first nodes respectively connected with a calculating machine, and two or more second nodes respectively connected with a storage subsystem,

Said first access request has the first destination ID which shows either of said two or more first nodes,

Said switch is a system characterized by changing said first destination ID into the second destination ID which shows either of said two or more storage subsystems, and transmitting through the second node connected with the storage subsystem which this second destination ID shows the second access request which has this second destination ID.

[Claim 18]

It is the switch connected to a calculating machine and two or more storage systems,

Two or more nodes respectively connected to a calculating machine, the first storage system, or the second storage system,

It has switching equipment connected with said two or more nodes,

One of nodes receives from a computer the access request of addressing to the second logic storage region which consists of a first logic storage region which said first storage system has, and a first logic storage region which said second storage system has,

Said switching equipment is a switch characterized by addressing according to the access request which received to either the first [said] storage system which has the first logic storage region which stores the data for access, or said second storage system, and transmitting an access request.

[Claim 19]

In a switch according to claim 18,

Said switch is a switch characterized by to change into the second identification information which shows either of the first logic storage regions where the first logic storage region which said first storage system has, or said second storage system has the first identification information which shows said second logic storage region included in the access request which received, to address the access request which has this second identification information to either said first storage system or said second storage system, and to transmit.

[Claim 20]

It is the system which memorizes the data accessed from a calculating machine,

The first storage system and the second storage system which have respectively the controller connected to a disk and a disk,

It has the switch connected to a calculating machine, said first storage system, and said second storage system,

Said switch is a system characterized by receiving the access request to the second logic storage region including the first logic storage region which is a storage region in the disk of said first storage system, and the first logic storage region which is a storage region in the disk of said second storage system from said calculating machine, and transmitting an access request to either said first storage system or the second storage system according to the access request which received.

[Claim 21]

It is a system according to claim 20,

It is the system characterized by transmitting the access request which the first identification information which shows said second logic storage region is contained in the access request which said switch receives, and said switch changes said first identification information into the second identification information which shows either the first logic storage region of said first storage system, or the first logic storage region of said second storage system, and has the second identification information to either said first storage system or said second storage system.

[Translation done.]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報**（Ａ）　(11)特許出願公開番号
特開2000－242434
（P2000－242434A）
(43)公開日　平成12年９月８日(2000.9.8)

(51)Int.Cl.⁷	識別記号	ＦＩ	テーマコード(参考)
Ｇ０６Ｆ　3/06	３０１	Ｇ０６Ｆ　3/06	３０１Ｇ
	５４０		５４０

審査請求　未請求　請求項の数20　ＯＬ　（全 24 頁）

(21)出願番号　特願平11－344260
(22)出願日　平成11年12月３日(1999.12.3)
(31)優先権主張番号　特願平10－364079
(32)優先日　平成10年12月22日(1998.12.22)
(33)優先権主張国　日本（ＪＰ）

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　松並　直人
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内
(72)発明者　大枝　高
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内
(74)代理人　100075096
弁理士　作田　康夫

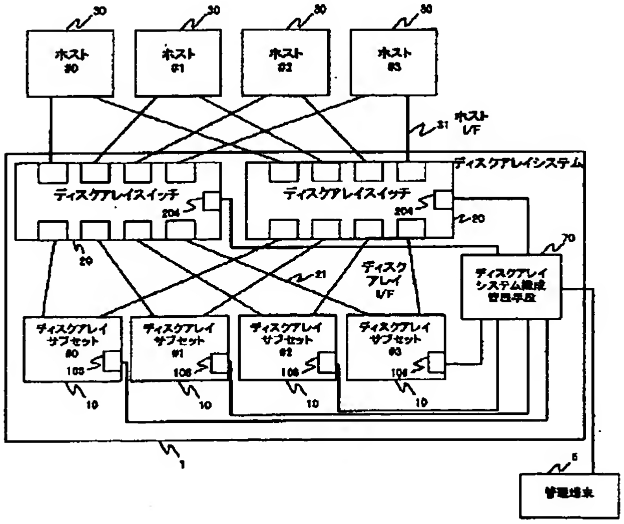
最終頁に続く

(54)【発明の名称】　記憶装置システム

(57)【要約】
【課題】計算機システムの規模、要求などに応じた記憶装置システムを構築でき、将来における記憶装置システムの拡張、信頼性の向上を容易に実現できるようにする。

【解決手段】記憶装置システム１は、データを保持する記憶装置とそれを制御する制御装置を有する複数のサブセット１０とサブセット１０とホスト３０との間に配置されるスイッチ装置２０を有する。スイッチ装置２０は、記憶装置システム１の構成を管理する管理情報を保持する管理テーブルを有し、管理情報に従ってホスト３０が出力するフレーム情報に含まれるアドレス情報を変換してフレーム情報をサブセット１０に振り分ける。

図１



【特許請求の範囲】

【請求項１】データを保持する記憶媒体を有する記憶装置と該記憶装置を制御する制御装置とを有する複数の記憶装置サブシステムと、前記複数の記憶装置サブシステムに保持されるデータを使用する計算機に接続され、記憶装置システムの構成情報を格納した構成管理テーブルと、前記計算機から送られてくるフレームに応答して、該フレームを解析し、前記構成管理テーブルに保持された構成情報に基づいて前記フレームを変換するフレーム変換手段とを有する第１のインタフェースノードと、各々が前記記憶装置サブシステムのいずれか１つに接続された複数の第２のインタフェースノードと、前記第１のインタフェースノード及び前記複数の第２のインタフェースノードが接続され、前記第１のインタフェースノードと前記複数の第２のインタフェースノードとの間で前記フレームの転送を行う転送手段とを有することを特徴とする記憶装置システム。

【請求項２】前記第１のインタフェースノードは、前記フレームに前記第２のインタフェースノードのノードアドレス情報を付加して出力するパケット生成手段を有し、前記転送手段は、前記ノードアドレス情報に基づいて前記第１のインタフェースノードと前記複数の第２のインタフェースノードとの間で前記フレームの転送を行うことを特徴とする請求項１記載の記憶装置システム。

【請求項３】前記フレームは、転送元及び転送先を指定する識別子を保持するヘッダ部と、転送される実体データを保持するデータ実体部とを有し、前記変換手段は、前記構成情報に基づき前記ヘッダ部に保持された転送先の識別子を変換することを特徴とする請求項１記載の記憶装置システム。

【請求項４】前記フレームは、前記データ実体部に、前記計算機により認識されている第１の論理アドレス情報を含み、前記変換手段は、前記構成管理テーブルに保持された前記構成情報に基づいて、前記第１の論理アドレス情報を、該フレームの転送先となる記憶装置サブシステム内で管理される第２の論理アドレスに変換することを特徴とする請求項３記載の記憶装置システム。

【請求項５】前記記憶装置システムは、さらに、前記転送手段に接続し、オペレータから記憶装置システムの構成を定義する構成情報の入力を受け付け、該入力に応答して、各ノードの前記構成管理テーブルに前記構成情報を設定する管理プロセッサを有することを特徴とする請求項１記載の記憶装置システム。

【請求項６】前記構成情報は、前記計算機から前記複数の記憶装置サブシステムへのアクセスを制限する情報を含むことを特徴とする請求項５記載の記憶装置システム。

【請求項７】前記第１のインタフェースノードは、前記計算機から転送されてくるデータの書き込みを指示するライトコマンドフレームに応答して、該ライトコマンドフレーム及びそれに続くデータフレームについてそれらの複製を生成し、前記ライトコマンドフレーム及びそれに続くデータフレームが少なくとも２つの記憶装置サブシステムに送られるよう、各々のフレームに異なるノードアドレス情報を付加して前記転送手段に転送することを特徴とする請求項２記載の記憶装置システム。

【請求項８】前記第１のインタフェースノードは、前記計算機から転送されてくるデータのリードを指示するリードコマンドフレームに応答して、該リードコマンドフレームの複製を生成し、前記少なくとも２つの記憶装置サブシステムに前記リードコマンドフレームが送られるように、各々のリードコマンドフレームに異なるノードアドレス情報を付加して前記転送手段に転送することを特徴とする請求項７記載の記憶装置サブシステム。

【請求項９】前記第１のインタフェースノードは、前記リードコマンドフレームに応答して前記少なくとも２つの記憶装置サブシステムから転送されてくるデータフレームを受信し、その一方を選択して前記計算機に転送することを特徴とする請求項８記載の記憶装置システム。

【請求項１０】前記第１のインタフェースノードは、前記計算機から転送されてくるデータのリードを指示するリードコマンドフレームに応答して、前記少なくとも２つの記憶装置サブシステムのうち予め定められた一の記憶装置サブシステムに接続する第２のインタフェースノードのノードアドレス情報を前記リードコマンドフレームに付加して前記転送手段に転送することを特徴とする請求項７記載の記憶装置サブシステム。

【請求項１１】データを保持する記憶媒体を有する記憶装置、及び該記憶装置を制御する制御装置とを有する複数の記憶装置サブシステムと、前記記憶装置に格納されたデータを利用する計算機との間に接続されるスイッチ装置であって、前記計算機に接続され、記憶装置システムの構成情報を格納した構成管理テーブルと、前記計算機から送られてくるフレームに応答して、該フレームを解析し、前記構成管理テーブルに保持された前記構成情報に基づいて前記フレームを変換する変換手段と、各々が前記記憶装置サブシステムのいずれかに接続された複数の第２のインタフェースノードと、前記第１のインタフェースノード及び前記複数の第２のインタフェースノードが接続され、前記第１のインタフェースノードと前記複数の第２のインタフェースノードとの間で前記フレームの転送を行う転送手段とを有すること特徴とするスイッチ装置。

【請求項１２】前記第１のインタフェースノードが、前記フレームに前記第２のインタフェースノードのノードアドレス情報を付加して出力するパケット生成手段を有し、前記転送手段は、前記ノードアドレス情報に基づいて前記第１のインタフェースノードと前記複数の第２のインタフェースノードとの間で前記フレームの転送を行うことを特徴とする請求項１１記載のスイッチ装置。

【請求項１３】前記フレームは、転送元及び転送先を指定する識別子を保持するヘッダ部と、転送される実体データを保持するデータ実体部とを有し、前記変換手段は、前記構成情報に基づき前記ヘッダ部に保持された転送先の識別子を変換することを特徴とする請求項１１記載のスイッチ装置。

【請求項１４】前記フレームは、前記データ実体部に、前記計算機により認識されている前記データの格納先を示す第１の論理アドレス情報を含み、前記変換手段は、前記構成管理テーブルに保持された前記構成情報に基づいて、前記第１の論理アドレス情報を、該フレームの転送先となる記憶装置サブシステム内で管理される第２の論理アドレスに変換することを特徴とする請求項１３記載のスイッチ装置。

【請求項１５】前記スイッチ装置は、さらに、前記転送手段に接続し、オペレータから該スイッチ装置及び前記複数の記憶装置サブシステムを含んで構成される記憶装置システムの構成を定義する構成情報の入力を受け付け、該入力に応答して、各ノードの構成管理テーブルに前記構成情報を設定する管理プロセッサを有することを特徴とする請求項１１記載のスイッチ装置。

【請求項１６】前記第１のインタフェースノードは、前記計算機から転送されてくるデータの書き込みを指示するライトコマンドフレームに応答して、該ライトコマンドフレーム及びそれに続くデータフレームについてそれらの複製を生成し、前記ライトコマンドフレーム及びそれに続くデータフレームが少なくとも２つの記憶装置サブシステムに送られるよう、各々のフレームに異なるノードアドレス情報を付加して前記転送手段に転送することを特徴とする請求項１２記載のスイッチ装置。

【請求項１７】前記第１のインタフェースノードは、前記計算機から転送されてくるデータのリードを指示するリードコマンドフレームに応答して、該リードコマンドフレームの複製を生成し、前記少なくとも２つの記憶装置サブシステムに前記リードコマンドフレームが送られるように、各々のリードコマンドフレームに異なるノードアドレス情報を付加して前記転送手段に転送することを特徴とする請求項１６記載のスイッチ装置。

【請求項１８】前記第１のインタフェースノードは、前記リードコマンドフレームに応答して前記少なくとも２つの記憶装置サブシステムから転送されてくるデータフレームを受信し、その一方を選択して前記計算機に転送することを特徴とする請求項１７記載のスイッチ装置。

【請求項１９】前記第１のインタフェースノードは、前記計算機から転送されてくるデータのリードを指示するリードコマンドフレームに応答して、前記少なくとも２つの記憶装置サブシステムのうち予め定められた一の記憶装置サブシステムに接続する第２のインタフェースノードのノードアドレス情報を前記リードコマンドフレームに付加して前記転送手段に転送することを特徴とする請求項１６記載のスイッチ装置。

【請求項２０】データを保持する記憶媒体を有する記憶装置と、該記憶装置を制御する制御装置とを有する複数の記憶装置サブシステムと、前記複数の記憶装置サブシステムに保持されるデータを使用する計算機に接続された第１のインタフェースノードと、各々が前記記憶装置サブシステムのいずれか１つに接続された複数の第２のインタフェースノードと、前記第１のインタフェースノード及び前記複数の第２のインタフェースノードが接続され、前記第１のインタフェースノードと前記複数の第２のインタフェースノードとの間でフレームの転送を行う転送手段と、前記転送手段に接続し、オペレータにより入力された記憶装置システムの構成を定義する構成情報を保持する管理テーブルを備えて前記構成情報に基づいて該記憶装置システムの構成を管理する管理プロセッサとを有することを特徴とする記憶装置システム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、複数のディスク装置を制御するディスク制御システムの実現方法に関し、特に、ディスク制御システムの高速化、低コスト化、コストパフォーマンスの向上の方法に関する。

【０００２】

【従来の技術】計算機システムに用いられる記憶装置システムとして、複数のディスク装置を制御するディスクアレイシステムがある。ディスクアレイシステムについては、例えば、“A Case for Redundant Arrays of Inexpensive Disks (RAID)”; InProc. ACM SIGMOD, June 1988（カリフォルニア大学バークレー校発行）に開示されている。ディスクアレイは、複数のディスク装置を並列に動作させることで、ディスク装置を単体で用いた記憶装置システムに比べ高速化を実現する技術である。

【０００３】複数のディスクアレイシステムを、複数のホストと相互に接続する方法として、ファイバチャネル（Fibre Channel）のFabricを使用した方法がある。この方法を適用した計算機システムの例が、日経エレクトロニクス1995.7.3（no.639）「シリアルSCSIがいよいよ市場へ」P.79 図３に示されている。ここに開示される計算機システムでは、複数のホストコンピュータ（以下では単にホストと呼ぶ）と複数のディスクアレイシステムが、それぞれ、ファイバチャネルを介してファブリック装置に接続される。ファブリック装置は、ファイバチャネルのスイッチであり、ファブリック装置に接続する任意の装置間の転送路の接続を行う。ファブリック装置はファイバチャネルのパケットである「フレーム」の転送に対し透過であり、ホストとディスクアレイシステムは、互いにファブリック装置を意識することなく２点間で通信を行う。

【０００４】

【発明が解決しようとする課題】従来のディスクアレイ

システムでは、大容量化のためディスク装置の台数を増やし、高性能化のため台数に見合った性能を有するコントローラを実現しようとすると、コントローラの内部バスの性能限界や、転送制御を行うプロセッサの性能限界が顕在化する。このような問題に対処するために、内部バスを拡張し、プロセッサ数を増加することが行われている。しかし、このような対処の仕方は、多数のバス制御によるコントローラ構成の複雑化や、プロセッサ間の共有データの排他制御等による制御ソフトの複雑化とオーバヘッドの増加を招く。このため、コストを非常に上昇させるとともに、性能は頭打ちになり、その結果、コストパフォーマンスが悪化する。また、このような装置は、大規模なシステムでは、そのコストに見合った性能が実現できるものの、規模がそれほど大きくないシステムには見合わない、拡張性が制限される、開発期間の増大と開発コストの上昇を招くといった課題がある。

【０００５】複数のディスクアレイシステムを並べファブリック装置で相互接続することによって、システム全体としての大容量化、高性能化を行うことが可能である。しかし、この方法では、ディスクアレイシステム間に関連性は全くなく、特定のディスクアレイシステムにアクセスが集中したとしてもそれを他の装置に分散することができないので、実使用上の高性能化が実現できない。また、ホストから見た論理的なディスク装置（論理ユニットと呼ぶ）の容量は、１台のディスクアレイシステムの容量に制限されるので、論理ユニットの大容量化は実現できない。

【０００６】ディスクアレイシステム全体を高信頼化しようとした際に、ホストが備えているミラーリング機能を用いて２台のディスクアレイシステムによるミラー構成を実現することができるが、ホストによるミラーリングのための制御オーバヘッドが発生し、システム性能が制限されるという課題がある。また、多数のディスクアレイシステムがシステム内に個別に存在すると、システム管理者が管理するための負荷が増加する。このため、多数の保守人員、複数台分の保守費用が必要になる等、管理コストが増加する。さらに、複数のディスクアレイシステム、ファブリック装置は、それぞれ独立した装置であるので、各種設定は、それぞれの装置毎に異なる方法で実施する必要がある。このため、管理者のトレーニングや、操作時間の増大にともない運用コストが増大する。

【０００７】本発明の目的は、これら従来技術における課題を解決し、計算機システムの規模、要求などに応じた記憶装置システムを構築でき、将来における記憶装置システムの拡張、信頼性の向上などに容易に対応することのできる記憶装置システムを実現することにある。

【０００８】

【課題を解決するための手段】本発明の記憶装置システムは、データを保持する記憶媒体を有する記憶装置と、この記憶装置を制御する制御装置とを有する複数の記憶装置サブシステム、複数の記憶装置サブシステムに保持されるデータを使用する計算機に接続された第１のインタフェースノード、各々が記憶装置サブシステムのいずれかに接続された複数の第２のインタフェースノード、及び第１のインタフェースノード及び複数の第２のインタフェースノードが接続され、第１のインタフェースノードと複数の第２のインタフェースノードとの間でフレームの転送を行う転送手段を有する。

【０００９】好ましくは、第１のインタフェースノードは、記憶装置システムの構成情報を格納した構成管理テーブルと、計算機から送られてくるフレームに応答して、該フレームを解析し、構成管理テーブルに保持された構成情報に基づいてそのフレームの転送先に関する情報変換して転送手段に転送する。

【００１０】また、フレームの転送に際して、第１のインタフェースノードは、そのフレームを受け取るべきノードのノードアドレス情報をフレームに付加する。転送手段はフレームに付加されたノードアドレス情報に従ってフレームを転送する。第２のインタフェースノードは、転送手段から受け取ったフレームからノードアドレス情報を除いてフレームを再形成し、目的の記憶装置サブシステムに転送する。

【００１１】本発明のある態様において、記憶装置システムは、転送手段に接続する管理プロセッサを有する。管理プロセッサは、オペレータからの指示に従って、構成管理テーブルに構成情報を設定する。構成情報には、計算機からのアクセスを制限する情報が含まれる。

【００１２】

【発明の実施の形態】［第１実施形態］図１は、本発明が適用されたディスクアレイシステムを用いたコンピュータシステムの一実施形態における構成図である。

【００１３】1はディスクアレイシステム、３０はディスクアレイシステムが接続されるホストコンピュータ（ホスト）である。ディスクアレイシステム1は、ディスクアレイサブセット１０、ディスクアレイスイッチ２０、ディスクアレイシステム全体の設定管理を行うディスクアレイシステム構成管理手段７０、ディスクアレイスイッチ２０とディスクアレイシステム構成管理手段７０との間、およびディスクアレイサブセット１０ディスクアレイシステム構成管理手段７０との間の通信インタフェース（通信Ｉ／Ｆ）８０を有する。ホスト３０とディスクアレイシステム1とは、ホストインタフェース（ホストＩ／Ｆ）３１で接続されており、ホストＩ／Ｆ３１はディスクアレイシステム1のディスクアレイスイッチ２０に接続する。ディスクアレイシステム1の内部において、ディスクアレイスイッチ２０とディスクアレイサブセット１０は、ディスクアレイインタフェース（ディスクアレイＩ／Ｆ２１）で接続される。

【００１４】ホスト３０、ディスクアレイサブセット１

０は、図では、各々４台示されているが、この台数に関しては制限はなく任意である。ホスト３０とディスクアレイサブセット１０の台数が異なっても構わない。また、ディスクアレイスイッチ２０は、本実施形態では図示の通り二重化されている。各ホスト３０および各ディスクアレイサブセット１０は、それぞれ別々のホストＩ／Ｆ３１、ディスクアレイＩ／Ｆ２１で二重化されたディスクアレイスイッチ２０の双方に接続されている。これは、一方のディスクアレイスイッチ２０、ホストＩ／Ｆ３１、あるいはディスクアレイＩ／Ｆ２１が故障しても他方を使用することでホスト３０からディスクアレイシステム1へのアクセスを可能とし、高い可用性を実現するためである。しかし、このような二重化は必ずしも必須ではなく、システムに要求される信頼性レベルに応じて選択可能である。

【００１５】図２は、ディスクアレイサブセット１０の一構成例を示す構成図である。１０１は上位システム（ホスト１０）からのコマンドを解釈してキャッシュヒットミス判定を実施し、上位システムとキャッシュ間のデータ転送を制御する上位アダプタ、１０２はディスクデータアクセス高速化のためのキャッシュ、および、マルチプロセッサ間の共有データを格納する共有メモリ（以下キャッシュ・共有メモリと呼ぶ）、１０４はディスクアレイサブセット10内に格納される複数のディスクユニットである。１０３はディスクユニット１０４を制御し、ディスクユニット１０４とキャッシュ間のデータ転送を制御する下位アダプタである。１０６はディスクアレイサブセット構成管理手段であり、ディスクアレイシステム1全体を管理するディスクアレイシステム構成管理手段７０と通信Ｉ／Ｆ８０を介して通信し、構成パラメータの設定や、障害情報の通報等の管理を行う。

【００１６】上位アダプタ１０１、キャッシュ・共有メモリ１０２、下位アダプタ１０３はそれぞれ二重化されている。この理由は上記ディスクアレイスイッチ２０の二重化と同様、高可用性を実現するためであり必須ではない。また、各ディスクユニット１０４は、二重化された下位アダプタ１０３のいずれからも制御可能である。本実施形態では、低コスト化の観点から同一のメモリ手段をキャッシュと共有メモリに共用しているが、これらは勿論分離することも可能である。

【００１７】上位アダプタ１０１は、上位アダプタ１０１の制御を実行する上位ＭＰＵ１０１０、上位システム、すなわちディスクアレイスイッチ２０との接続Ｉ／ＦであるディスクアレイＩ／Ｆ２１を制御するディスクアレイＩ／Ｆコントローラ１０１１、キャッシュ・共有メモリ１０２と上位ＭＰＵ１０１０とディスクアレイＩ／Ｆコントローラ１０１１との間の通信、データ転送を行う上位バス１０１２を含む。

【００１８】図では各上位アダプタ１０１毎に1台のディスクアレイＩ／Ｆコントローラ１０１１が示されているが、１つの上位アダプタに対し、複数のディスクアレイＩ／Ｆコントローラ１０１１を設けてもよい。

【００１９】下位アダプタ１０３は、下位アダプタ１０３の制御を実行する下位ＭＰＵ１０３０、ディスク１０４とのインタフェースであるディスクＩ／Ｆを制御するディスクＩ／Ｆコントローラ１０３１、キャッシュ・共有メモリ１０２と下位ＭＰＵ１０３０とディスクＩ／Ｆコントローラ１０３１との間の通信、データ転送を行う下位バス１０３２を含む。

【００２０】図では各下位アダプタ１０３毎に4台のディスクＩ／Ｆコントローラ１０３１が示されているが、その数は任意であり、ディスクアレイの構成や、接続するディスク台数に応じて変更可能である。

【００２１】図３は、ディスクアレイスイッチ２０の一構成例を示す構成図である。２００はディスクアレイスイッチ全体の制御および管理を行うプロセッサである管理プロセッサ（ＭＰ）、２０１はｎ×ｎの相互スイッチ経路を構成するクロスバスイッチ、２０２はディスクアレイＩ／Ｆ２１毎に設けられるディスクアレイＩ／Ｆノード、２０３はホストＩ／Ｆ３１毎に設けられるホストＩ／Ｆノード、２０４はディスクアレイシステム構成管理手段７０との間の通信を行う通信コントローラである。２０２０はディスクアレイＩ／Ｆノード２０２とクロスバスイッチ２０１を接続するパス、２０３０はホストＩ／Ｆノード２０３とクロスバスイッチ２０１を接続するパス、２０４０は他のディスクアレイスイッチ２０と接続し、クラスタを構成するためのクラスタ間Ｉ／Ｆ、２０５０はＭＰ２００とクロスバスイッチ２０１を接続するためのパスである。

【００２２】図４はクロスバスイッチ２０１の構造を示す構成図である。２０１０はクロスバスイッチ２０１に接続するパス２０２０、２０３０、２０５０、およびクラスタ間Ｉ／Ｆ２０４０を接続するポートであるスイッチングポート（ＳＷＰ）である。ＳＷＰ２０１０はすべて同一の構造を有し、あるＳＷＰから他のＳＷＰへの転送経路のスイッチング制御を行う。図では１つのＳＷＰについてのみ転送経路を示しているが、すべてのＳＷＰ間で同様の転送経路が存在する。

【００２３】図５は、ホストＩ／Ｆノード２０３の一構成例を示す構成図である。本実施形態では、具体的に説明をするためにホストＩ／Ｆ３１とディスクアレイＩ／Ｆ２１の両方にファイバチャネルを使用するものと仮定する。もちろんホストＩ／Ｆ３１とディスクアレイＩ／Ｆ２１として、ファイバチャネル以外のインタフェースを適用することも可能である。ホストＩ／Ｆノード２０３とディスクアレイＩ／Ｆノード２０２の両方に同一のインタフェースを使用することで、両者を同一構造にできる。本実施形態においては、ディスクアレイＩ／Ｆノード２０２も図に示すホストＩ／Ｆノード２０３と同様に構成される。以下では、ホストＩ／Ｆノード２０３を

例に説明を行う。

【００２４】２０２１は受信したファイバチャネルフレーム（以下単にフレームと呼ぶ）をどのノードに転送するかを検索する検索プロセッサ（ＳＰ）、２０２２はホスト３０（ディスクアレイＩ／Ｆノード２０２の場合は、ディスクアレイサブセット１０）との間でフレームを送受信するインタフェースコントローラ（ＩＣ）、２０２２はＩＣ２０２３が受信したフレームに対しＳＰ２０２１が検索した結果に基づいて変換を施すスイッチングコントローラ（ＳＣ）、２０２４はＳＣ２０２１が変換したフレームを他のノードに転送するためにクロスバスイッチ２０１を通過できる形式にパケット化するパケット生成部（ＳＰＧ）、２０２５は受信したフレームを一時的に格納するフレームバッファ（ＦＢ）、２０２６は一つのホストからのディスクアレイアクセス要求コマンド（以下単にコマンドと呼ぶ）に対応した複数のフレーム列であるエクスチェンジ（Exchange）を識別するためのエクスチェンジ番号を管理するエクスチェンジテーブル（ＥＴ）、２０２７は複数のディスクアレイサブセット１０の構成情報を格納するディスクアレイ構成管理テーブル（ＤＣＴ）である。

【００２５】ディスクアレイスイッチ２０の各構成部は、すべてハードウェアロジックで構成されることが性能上望ましい。しかし、求められる性能を満足できるならば、汎用プロセッサを用いたプログラム制御によりＳＰ２０２１やＳＣ２０２２の機能を実現することも可能である。

【００２６】各ディスクアレイサブセット１０は、各々が有するディスクユニット１０４を１または複数の論理的なディスクユニットとして管理している。この論理的なディスクユニットを論理ユニット（ＬＵ）と呼ぶ。ＬＵは、物理的なディスクユニット１０４と１対１で対応する必要はなく、１台のディスクユニット１０４に複数のＬＵが構成され、あるいは、複数のディスクユニット１０４で１つのＬＵが構成されても構わない。

【００２７】ディスクアレイサブセット１０の外部から見た場合、１つのＬＵは、１台のディスク装置として認識される。本実施形態では、ディスクアレイスイッチ２０によりさらに論理的なＬＵが構成され、ホスト３０は、このＬＵに対してアクセスするように動作する。本明細書では、１つのＬＵでホスト３０から認識される１つのＬＵが構成される場合、ホスト３０により認識されるＬＵを独立ＬＵ（ＩＬＵ）、複数のＬＵでホスト３０から認識される１つのＬＵが構成される場合、ホスト３０により認識されるＬＵを統合ＬＵ（ＣＬＵ）と呼ぶ。

【００２８】図１２に、４つのディスクアレイサブセットのＬＵで１つの統合ＬＵが構成される場合における各階層間でのアドレス空間の対応関係を示す。図において、１０００は、一例として、ホスト“＃２”からみたディスクアレイシステム１の１つの統合ＬＵにおけるアドレス空間、１１００は、ディスクアレイサブセット１０のＬＵのアドレス空間、１２００はディスクユニット１０４（ここでは、ディスクアレイサブセット“＃０”についてのみ図示されている）のアドレス空間を示している。

【００２９】各ディスクアレイサブセット１０のＬＵは、ここでは、４台のディスクユニット１０４によりＲＡＩＤ５（Ｒｅｄｕｎｄａｎｔ Ａｒｒａｙｓ ｏｆ Ｉｎｅｘｐｅｎｓｉｖｅ Ｄｉｓｋｓ Ｌｅｖｅｌ ５）型ディスクアレイとして構成されるものとする。各ディスクアレイサブセット１０は、それぞれｎ０、ｎ１、ｎ２、ｎ３の容量を有するＬＵを持つ。ディスクアレイスイッチ２０は、これら４つのＬＵの持つアドレス空間を（ｎ０＋ｎ１＋ｎ２＋ｎ３）の容量を有するアドレス空間に統合し、ホスト３０から認識される統合ＬＵを実現する。

【００３０】本実施形態では、例えば、ホスト＃２が領域Ａ１００１をアクセスする場合、領域Ａ１００１を指定したアクセス要求は、ディスクアレイスイッチ２０によりディスクアレイサブセット＃０のＬＵの領域Ａ’１１０１をアクセスするための要求に変換されてディスクアレイサブセット＃０に転送される。ディスクアレイサブセット＃０は、領域Ａ’１１０１をさらに、ディスクユニット１０４上の領域Ａ”１２０１にマッピングしてアクセスを行う。アドレス空間１０００とアドレス空間１１００との間のマッピングは、ディスクアレイスイッチ２０が有するＤＣＴ２０７に保持された構成情報に基づき行われる。この処理の詳細については後述する。なお、ディスクアレイサブセット内におけるマッピングについては、既によく知られた技術であり、本明細書では詳細な説明については省略する。

【００３１】本実施形態において、ＤＣＴ２０７は、システム構成テーブルとサブセット構成テーブルを含む。図６は、システム構成テーブルの構成を、図７は、サブセット構成テーブルの構成を示す。

【００３２】図７に示すように、システム構成テーブル２０２７０は、ホストＬＵの構成を示す情報を保持するホストＬＵ構成テーブル20271、及びディスクアレイスイッチ２０のディスクアレイＩ／Ｆノード２０２とディスクアレイサブセット１０との接続関係を示すディスクアレイＩ／Ｆノード構成テーブル20272を有する。

【００３３】ホストＬＵ構成テーブル20271は、ホスト３０からみたＬＵごとに、そのＬＵを識別する番号であるHost-LU No.、ＬＵの属性を示すLU Type、CLU Class、及びCLU Stripe Size、ホストＬＵの状態を示す情報であるCondition、ホストＬＵを構成するディスクアレイサブセット１０のＬＵに関する情報であるＬＵ情報（LU Info.）を有する。

【００３４】LU Typeは、このホストＬＵがＣＬＵであるか、ＩＬＵであるかといったＬＵの種類を示す情報で

ある。CLU Classは、LU TypeによりこのホストＬＵがＣＬＵであることが示される場合に、そのクラスが“Joined”、“mirrored”、及び“Striped”のいずれであるかを示す情報である。“Joined”は、図１１により説明したように、いくつかのＬＵを連結して１つの大きな記憶空間を持つＣＬＵが構成されていることを示す。“Mirrored”は、第６実施形態として後述するように、２つのＬＵにより二重化されたＬＵであることを示す。“Striped”は、第７実施形態として後述するように、複数のＬＵで構成され、データがこれら複数のＬＵに分散して格納されたＬＵであることを示す。CLU Stripe Sizeは、CLU Classにより「Striped」であることが示される場合に、ストライピングサイズ（データの分散の単位となるブロックのサイズ）を示す。

【００３５】Conditionにより示される状態には、“Normal”、“Warning”、“Fault”、及び“Not Defined”の４種類がある。“Normal”はこのホストＬＵが正常な状態であることを示す。“Warning”は、このホストＬＵを構成するＬＵに対応するいずれかのディスクユニットに障害が発生している等の理由により縮退運転が行われていることを示す。“Fault”は、ディスクアレイサブセット１０の故障などによりこのホストＬＵを運転することができないことを示す。“Not Defined”は、対応するHost-LU No.のホストＬＵが定義されていないことを示す。

【００３６】LU Infoは、このホストＬＵを構成するＬＵについて、そのＬＵが属するディスクアレイサブセット１０を特定する情報、ディスクアレイサブセット内でのＬＵＮ、及びそのサイズを示す情報を含む。ホストＬＵがＩＬＵの場合には、唯一のＬＵに関する情報が登録される。ホストＬＵがＣＬＵの場合には、それを構成する全てのＬＵについて、それぞれのＬＵに関する情報が登録される。例えば、図において、Host-LU No.が“０”であるHost-LUは、ディスクアレイサブセット“＃０”のＬＵＮ“０”、ディスクアレイサブセット“＃１”のＬＵＮ“０”、ディスクアレイサブセット“＃２”のＬＵＮ“０”、ディスクアレイサブセット“＃３”のＬＵＮ“０”の4つのＬＵから構成されるＣＬＵであり、そのＣＬＵクラスが“Joined”であるＣＬＵであることが分かる。

【００３７】ディスクアレイＩ／Ｆノード構成テーブル20272は、ディスクアレイＩ／Ｆ２１が接続するディスクアレイサブセット１０のポートごとに、どのディスクアレイスイッチ２０のディスクアレイＩ／Ｆノード２０２が接続されるかを示す情報を保持する。

【００３８】具体的には、ディスクアレイサブセット１０を特定するSubset No.、ポートを特定するSubset Port No.、そのポートに接続するディスクアレイスイッチ２０を特定するSwitch No.、及びそのディスクアレイスイッチ２０のディスクアレイＩ／Ｆノード２０２を特定するI/F Node No.を有する。ディスクアレイサブセット１０が複数のポートを備えている場合には、そのポート毎に情報が設定される。

【００３９】サブセット構成テーブルは、図７に示すように、各ディスクアレイサブセット１０に対応する複数のテーブル202720～202723を有する。各テーブルは、ディスクアレイサブセット１０内で構築されたＲＡＩＤグループの構成を示す情報を保持するＲＡＩＤグループ構成テーブル202730と、ディスクアレイサブセット１０内に構築されたＬＵの構成を示す情報を保持するＬＵ構成テーブル202740を含む。

【００４０】ＲＡＩＤグループ構成テーブル202730は、ＲＡＩＤグループに付加された番号を示すGroup No.、そのＲＡＩＤグループのレベルを示すLevel、そのＲＡＩＤグループを構成するディスクの数を示す情報であるDisks、そのＲＡＩＤグループがＲＡＩＤレベル０，５等のストライピングされた構成の場合、そのストライプサイズを示すStripe Sizeを情報として含む。例えば、図に示されるテーブルにおいて、ＲＡＩＤグループ“０”は、４台のディスクユニットにより構成されたＲＡＩＤグループであり、ＲＡＩＤレベルが５、ストライプサイズがＳ０である。

【００４１】ＬＵ構成テーブル202740は、ＬＵに付加された番号（ＬＵＮ）を示すLU No.、このＬＵがどのＲＡＩＤグループに構成されているのかを示すRAID Group、ＬＵの状態を示すCondition、このＬＵのサイズ（容量）を示すSize、このＬＵがディスクアレイサブセット１０のどのポートからアクセス可能なのかを示すPort、及びその代替となるポートを示すAlt. Portを情報として含む。Conditionで示される状態は、ホストＬＵについてのConditionと同様、“Normal”、“Warning”、“Fault”、“Not Defined”の４種類がある。Alt. Portに設定された情報により特定されるポートは、Portに設定された情報で特定されるポートに障害が発生したときに用いられるが、単に複数のポートから同一のＬＵをアクセスするために用いることもできる。

【００４２】図８は、ファイバチャネルにおけるフレームの構成図である。ファイバチャネルのフレーム４０は、フレームの先頭を示すＳＯＦ（Start Of Frame）４００、フレームヘッダ４０１、転送の実態データを格納する部位であるフレームペイロード４０２、３２ビットのエラー検出コードであるＣＲＣ（Cyclic RedundancyCheck）４０３、フレームの最後尾を示すＥＯＦ（End Of Frame）４０４を含む。フレームヘッダ４０１は、図９に示すような構造になっており、フレーム転送元のＩＤ（S_ID）、フレーム転送先のＩＤ（D_ID)、エクスチェンジの起動元、応答先が指定するそれぞれのエクスチェンジＩＤ（OX_ID、RX_ID）、エクスチェンジ中のフレームグループを指定するシーケンスのＩＤ（SEQ_ID）等が格納されている。

【００４３】本実施形態では、ホスト３０により発行されるフレームには、S_IDとしてホスト３０に割り当てられたＩＤが、また、D_IDとしてディスクアレイスイッチ２０のポートに割り当てられたＩＤが使用される。一つのホストコマンドに対し、１ペアのエクスチェンジＩＤ（OX_ID、RX_ID）が割り当てられる。複数のデータフレームを同一のエクスチェンジに対し発行する必要があるときは、その全データフレームに対して同一のSEQ_IDが割り当てられ、おのおのはシーケンスカウント（SEQ_CNT）で識別される。フレームペイロード４０２の最大長は２１１０バイトであり、フレーム種毎に格納される内容が異なる。例えば、後述するFCP_CMDフレームの場合、図１０に示すように、ＳＣＳＩのLogical Unit Number（ＬＵＮ）、Command Description Block（ＣＤＢ）等が格納される。ＣＤＢは、ディスク（ディスクアレイ）アクセスに必要なコマンドバイト、転送開始論理アドレス（ＬＢＡ）、転送長（ＬＥＮ）を含む。

【００４４】以下、本実施形態のディスクアレイシステムの動作を説明する。

【００４５】ディスクアレイシステムを使用するのに先立ち、ディスクアレイスイッチ２０に対して、ディスクアレイサブセット１０の構成情報を設定する必要がある。システム管理者は、管理端末５からディスクアレイシステム構成手段７０を介して、すべてのディスクアレイサブセット１０およびディスクアレイスイッチ２０の構成設定情報を獲得する。管理者は、管理端末５から所望のシステム構成になるよう論理ユニットの構成設定、RAIDレベルの設定、障害発生時の交代パスの設定等、各種設定に必要な設定情報を入力する。ディスクアレイシステム構成管理手段７０は、その設定情報を受け、各ディスクアレイサブセット１０およびディスクアレイスイッチ２０に設定情報を転送する。なお、管理端末５における設定情報の入力については第５実施形態にて別途説明する。

【００４６】ディスクアレイスイッチ２０では、通信コントローラ２０４が設定情報を獲得し、ＭＰ２００により各ディスクアレイサブセット１０のアドレス空間情報等の構成情報が設定される。ＭＰ２００は、クロスバスイッチ２０１経由で各ホストＩ／Ｆノード２０３およびディスクアレイＩ／Ｆノード２０２に、ディスクアレイサブセット１０の構成情報を配信する。

【００４７】各ノード２０３、および２０２はこの情報を受信すると、ＳＰ２０２１により構成情報をＤＣＴ２０２７に格納する。ディスクアレイサブセット１０では、ディスクアレイサブセット構成管理手段１０６が、設定情報を獲得し、共有メモリ１０２に格納する。各上位ＭＰＵ１０１０および下位ＭＰＵ１０３０は、共有メモリ１０２上の設定情報を参照し、各々の構成管理を実施する。

【００４８】以下では、ホスト“＃２”がディスクアレイシステム1に対し、リードコマンドを発行した場合の動作を説明する。図１１に、ホストからのリード動作時にファイバチャネルを通して転送されるフレームのシーケンスを示す模式図を、図１３にこのときのディスクアレイスイッチのホストＩ／Ｆノード２０３における動作のフローチャートを示す。

【００４９】なお、以下の説明では、ホスト“＃２”が、図１２における記憶領域Ａ１００１をアクセスすることを仮定する。記憶領域Ａ１００１に対応する実際の記憶領域Ａ″は、ディスクアレイサブセット“＃０”のＬＵＮ＝０のＬＵを構成するディスクユニット＃２のアドレス空間内に存在するものとする。また、アドレス空間１０００を構成するＬＵを定義しているホストＬＵ構成テーブル20271のLU Typeには「ＣＬＵ」が、CLU Classには「Joined」が設定されているものとする。

【００５０】データのリード時、ホスト３０は、リードコマンドを格納したコマンドフレーム「FCP_CMD」をディスクアレイスイッチ２０に発行する（図１１矢印（ａ））。ディスクアレイスイッチ２０のホストＩ／Ｆノード“＃２”は、ＩＣ２０２３によりホストＩ／Ｆ３１経由でコマンドフレーム「FCP_CMD」を受信する（ステップ20001）。ＩＣ２０２３は、ＳＣ２０２２にコマンドフレームを転送する。ＳＣ２０２２は、受け取ったコマンドフレームを一旦ＦＢ２０２５に格納する。この際、ＳＣ２０２２は、コマンドフレームのＣＲＣを計算し、受信情報が正しいことを検査する。ＣＲＣの検査に誤りがあれば、ＳＣ２０２２は、その旨をＩＣ２０２３に通知する。ＩＣ２０２３は、誤りの通知をＳＣ２０２２から受けると、ホストＩ／Ｆ３１を介してホスト３０にＣＲＣエラーを報告する。（ステップ20002）。

【００５１】ＣＲＣが正しい場合、ＳＣ２０２２は、ＦＢ２０２５に保持したフレームをリードし、それがコマンドフレームであることを認識してフレームヘッダ４０１を解析する（ステップ20003）。そして、ＳＣ２０２２は、ＳＰ２０２１に指示し、S_ID、D_ID、OX_ID等のエクスチェンジ情報をＥＴ２０２６に登録する（ステップ20004）。

【００５２】次に、ＳＣ２０２２は、フレームペイロード４０２を解析し、ホスト３０により指定されたＬＵＮおよびＣＤＢを取得する（ステップ20005）。ＳＰ２０２１は、ＳＣ２０２２の指示により、ＤＣＴ２０２７を検索し、ディスクアレイサブセット１０の構成情報を得る。具体的には、ＳＰ２０２１は、ホストＬＵ構成テーブル20271を検索し、受信したフレームペイロード４０２に格納されたＬＵＮと一致するHost-LU No.を有する情報を見つける。ＳＰ２０２１は、LU Type、CLU Classに設定された情報からホストＬＵの構成を認識し、LU Info.に保持されている情報に基づきアクセスすべきディスクサブセット１０とその中のＬＵのＬＵＮ、及びこのＬＵ内でのＬＢＡを判別する。次に、ＳＰ２０２１は、

サブセット構成テーブル202720のＬＵ構成テーブル202740を参照し、目的のディスクアレイサブセット１０の接続ポートを確認し、ディスクアレイＩ／Ｆノード構成テーブル20272からそのポートに接続するディスクアレイＩ／Ｆノード２０２のノードNo.を得る。ＳＰ２０２１は、このようにして得たディスクアレイサブセット１０を識別する番号、ＬＵＮ、ＬＢＡ等の変換情報をＳＣ２０２２に報告する。（ステップ20006）。

【００５３】次に、ＳＣ２０２２は、獲得した変換情報を使用しフレームペイロード４０２のＬＵＮとＣＤＢのなかのＬＢＡを変換する。また、フレームヘッダ４０１のD_IDを対応するディスクアレイサブセット１０のホストＩ／Ｆコントローラ１０１１のD_IDに変換する。なお、この時点ではS_IDは書き換えない（ステップ２０００７）。

【００５４】ＳＣ２０２２は、変換後のコマンドフレームと、対象ディスクアレイサブセット１０に接続するディスクアレイＩ／Ｆノード番号を、ＳＰＧ２０２４に転送する。ＳＰＧ２０２４は、受け取った変換後のコマンドフレームに対し、図１４に示すような簡単な拡張ヘッダ６０１を付加したパケットを生成する。このパケットをスイッチングパケット（Ｓ Ｐａｃｋｅｔ）６０と呼ぶ。S Packet６０の拡張ヘッダ６０１には、転送元（自ノード）番号、転送先ノード番号、及び転送長が付加含まれる。ＳＰＧ２０２４は、生成したS Packet６０をクロスバスイッチ２０１に送信する（ステップ20008）。

【００５５】クロスバスイッチ２０１は、ホストＩ／Ｆノード“＃２”と接続するＳＷＰ２０１０によりS Packet６０を受信する。ＳＷＰ２０１０は、S Packet６０の拡張ヘッダ６０１を参照し、転送先のノードが接続するＳＷＰへのスイッチ制御を行って経路を確立し、S Packet６０を転送先のディスクアレイＩ／Ｆノード２０２（ここでは、ディスクアレイＩ／Ｆノード“＃０”）に転送する。ＳＷＰ２０１０は、経路の確立をS Packet６０の受信の度に実施し、S Packet６０の転送が終了したら、その経路を解放する。ディスクアレイＩ／Ｆノード“＃０”では、ＳＰＧ２０２４がS Packet６０を受信し、拡張ヘッダ６０１を外してコマンドフレームの部分をＳＣ２０２２に渡す。

【００５６】ＳＣ２０２２は、受け取ったコマンドフレームのフレームヘッダのS_IDに自分のＩＤを書き込む。次にＳＣ２０２２は、ＳＰ２０２１に対し、コマンドフレームのS_ID、D_ID、OX_ID等のエクスチェンジ情報、及びフレーム転送元ホストＩ／Ｆノード番号をＥＴ２０２６に登録するよう指示し、ＩＣ２０２３にコマンドフレームを転送する。ＩＣ２０２３は、フレームヘッダ４０１の情報に従い、接続するディスクアレイサブセット１０（ここでは、ディスクアレイサブセット“＃０”）にコマンドフレームを転送する（図１１矢印（ｂ））。

【００５７】ディスクアレイサブセット“＃０”は、変換後のコマンドフレーム「FCP_CMD」をディスクアレイＩ／Ｆコントローラ１０１１で受信する。上位ＭＰＵ１０１０は、コマンドフレームのフレームペイロード４０２に格納されたＬＵＮとＣＤＢを取得し、指定された論理ユニットのＬＢＡからＬＥＮ長のデータをリードするコマンドであると認識する。

【００５８】上位ＭＰＵ１０１０は、共有メモリ１０２に格納されたキャッシュ管理情報を参照し、キャッシュヒットミス／ヒット判定を行う。ヒットすればキャッシュ１０２からデータ転送を実施する。ミスの場合、ディスクユニットからデータをリードする必要があるので、ＲＡＩＤ５の構成に基づくアドレス変換を実施し、キャッシュ空間を確保する。そして、ディスクユニット２からのリード処理に必要な処理情報を生成し、下位ＭＰＵ１０３０に処理を引き継ぐべく、共有メモリ１０２に処理情報を格納する。

【００５９】下位ＭＰＵ１０３０は、共有メモリ１０２に処理情報が格納されたことを契機に処理を開始する。下位ＭＰＵ１０３０は、適切なディスクＩ／Ｆコントローラ１０３１を特定し、ディスクユニット２へのリードコマンドを生成して、ディスクＩ／Ｆコントローラ１０３１にコマンドを発行する。ディスクＩ／Ｆコントローラ１０３１は、ディスクユニット2からリードしたデータをキャッシュ１０２の指定されたアドレスに格納して下位ＭＰＵ１０３０に終了報告を通知する。下位ＭＰＵ１０３０は、処理が正しく終了したことを上位ＭＰＵ１０１０に通知すべく共有メモリ１０２に処理終了情報を格納する。

【００６０】上位ＭＰＵ１０１０は、共有メモリ１０２に処理終了情報が格納されたことを契機に処理を再開し、ディスクアレイＩ／Ｆコントローラ１０１１にリードデータ準備完了を通知する。ディスクアレイＩ／Ｆコントローラ１０１１は、ディスクアレイスイッチ２０の当該ディスクアレイＩ／Ｆノード“＃０”に対し、ファイバチャネルにおけるデータ転送準備完了フレームである「FCP_XFER_RDY」を発行する（図１１矢印（ｃ））。

【００６１】ディスクアレイＩ／Ｆノード“＃０”では、データ転送準備完了フレーム「FCP_XFER_RDY」を受信すると、ＳＣ２０２２が、ディスクアレイサブセット２０から受信した応答先エクスチェンジＩＤ（RX_ID）を獲得し、S_ID、D_ID、OX_IDを指定して、ＳＰ２０２１に指示しＥＴ２０２６の当該エクスチェンジ情報にRX_IDを登録する。ＳＣ２０２２は、データ転送準備完了フレームの転送先（コマンドフレームの転送元）のホストＩ／Ｆノード番号を獲得する。ＳＣ２０２２は、このフレームのS_IDを無効化し、ＳＰＧ２０２４に転送する。ＳＰＧ２０２４は、先に述べたようにしてS Packetを生成し、クロスバスイッチ２０１経由で対象ホストＩ／Ｆノード“＃２”に転送する。

【００６２】ホストＩ／Ｆノード“＃２”では、ＳＰＧ

２０２４がデータ転送準備完了フレームのS Packetを受信すると、S Packetの拡張ヘッダを外し「FCP_XFER_RDY」を再生してＳＣ２０２２に渡す（ステップ20011）。ＳＣ２０２２は、ＳＰ２０２１に指示しＥＴ２０２６をサーチして該当するエクスチェンジを特定する（ステップ20012）。

【００６３】次に、ＳＣ２０２２は、フレームが「FCP_XFER_RDY」であるかどうか調べ（ステップ20013）、「FCP_XFER_EDY」であれば、ＥＴ２０２６の応答先エクスチェンジＩＤ（RX_ID）の更新をＳＰ２０２１に指示する。応答先エクスチェンジＩＤとしては、このフレームに付加されていた値が使用される（ステップ20014）。そして、ＳＣ２０２２は、フレームヘッダ４０１のS_ID、D_IDをホストＩ／Ｆノード２０３のＩＤとホスト３０のＩＤを用いた適切な値に変換する（ステップ20015）。これらの処理によりフレームヘッダ４０１は、ホスト“＃２”に対するフレームに変換される。ＩＣ２０２３は、ホスト“＃２”に対し、このデータ転送準備完了フレーム「FCP_XFER_RDY」を発行する（図１１の矢印（ｄ）：ステップ20016）。

【００６４】ディスクアレイサブセット“＃０”のディスクアレイＩ／Ｆコントローラ１０１１は、データ転送を行うため、データフレーム「FCP_DATA」を生成し、ディスクアレイスイッチ２０に転送する（図１１矢印（ｅ））。フレームペイロードの転送長には制限があるため、１フレームで転送できる最大のデータ長は２ＫＢである。データ長がこれを越える場合は、必要数だけデータフレームを生成し発行する。すべてのデータフレームには同一のSEQ_IDが割り当てられる。データフレームの発行は、同一のSEQ_IDに対し複数のフレームが生成されることを除き（すなわちSEQ_CNTが変化する）、データ転送準備完了フレームの場合と同様である。

【００６５】ディスクアレイスイッチ２０は、データ転送準備完了フレームの処理と同様に、データフレーム「FCP_DATA」のフレームヘッダ４０１の変換を実施する。ただし、データフレームの転送の場合、RX_IDが既に確立されているので、データ転送準備完了フレームの処理におけるステップ20014の処理はスキップされる。フレームヘッダ４０１の変換後、ディスクアレイスイッチ２０は、ホスト“＃２”にデータフレームを転送する（図１１矢印（ｆ））。

【００６６】次に、ディスクアレイサブセット“＃０”のディスクアレイＩ／Ｆコントローラ１０１１は、終了ステータス転送を行うため、ステータスフレーム「FCP_RSP」を生成し、ディスクアレイスイッチ２０に対し発行する（図１１矢印（ｇ））。ディスクアレイスイッチ２０では、データ転送準備完了フレームの処理と同様に、ＳＰＧ２０２４がS Packetから拡張ヘッダを外し「FCP_RSP」ステータスフレームを再現し（ステップ20021）、ＳＰ２０２１によりＥＴ２０２６を検索しエクスチェンジ情報を獲得する（ステップ20022）。ＳＣ２０２２は、その情報に基づきフレームを変換する（ステップの20023）。変換されたフレームは、ＩＣ２０２３によりホスト“＃２”に転送される（図１１矢印（ｈ）：ステップ20024）。最後にＳＰ２０２１は、ＥＴ２０２６からエクスチェンジ情報を削除する（ステップ20025）。

【００６７】以上のようにしてディスクアレイからのリード処理が行われる。ディスクアレイシステム１に対するライト処理についてもデータフレームの転送方向が逆転するのみで、上述したリード処理と同様の処理が行われる。

【００６８】図３に示したように、ディスクアレイスイッチ２０は、クロスバスイッチ２０１にクラスタ間Ｉ／Ｆ２０４０を備えている。図１に示したシステム構成では、クラスタ間Ｉ／Ｆ２０４０は使用されていない。本実施形態のディスクアレイスイッチ２０は、クラスタ間Ｉ／Ｆ２０４０を利用して図１５に示すように、他のディスクアレイスイッチと相互に接続されることができる。

【００６９】本実施形態におけるディスクアレイスイッチ２０単独では、ホスト３０とディスクアレイサブセット１０を合計８台までしか接続できないが、クラスタ間Ｉ／Ｆ２０４０を利用して複数のディスクアレイスイッチを相互接続し、接続できるホスト１０とディスクアレイの数を増やすことができる。例えば、図１５に示すシステムでは、４台のディスクアレイスイッチ２０を使ってホスト３０とディスクアレイサブセット１０を合計３２台まで接続でき、これらの間で相互にデータ転送が可能になる。

【００７０】このように、本実施形態では、ディスク容量や性能の必要性に合わせて、ディスクアレイサブセットやホストの接続台数を増加していくことができる。また、必要な転送帯域分のホストＩ／Ｆを用いてホストーディスクアレイシステム間を接続することができるので、容量、性能、接続台数の拡張性を大幅に向上させることができる。

【００７１】以上説明した実施形態によれば、１台のディスクアレイサブセットの性能が、内部のＭＰＵや内部バスで制限されたとしても、複数のディスクアレイサブセットを用いて、ディスクアレイスイッチによりホストとディスクアレイサブセット間を相互接続することができる。これにより、ディスクアレイシステムトータルとして高い性能を実現することができる。ディスクアレイサブセットの性能が比較的低いものであっても、複数のディスクアレイサブセットを用いることで高性能化を実現できる。したがって、低コストのディスクアレイサブセットをコンピュータシステムの規模に合わせて必要な台数だけ接続することができ、規模に応じた適切なコストでディスクアレイシステムを構築することが可能とな

る。

【００７２】また、ディスク容量の増大や性能の向上が必要になったときは、ディスクアレイサブセットを必要なだけ追加すればよい。さらに、複数のディスクアレイスイッチを用いて任意の数のホスト及びディスクアレイサブセットを接続できるので、容量、性能、接続台数のいずれをも大幅に向上させることができ、高い拡張性を有するシステムが実現できる。

【００７３】さらにまた、本実施形態によれば、ディスクアレイサブセットとして、従来のディスクアレイシステムそのものの縮小機を用いることができるので、既に開発した大規模な制御ソフトウェア資産をそのまま利用でき、開発コストの低減と開発期間の短縮を実現することができる。

【００７４】［第２実施形態］図１６は、本発明の第２の実施形態におけるコンピュータシステムの構成図である。本実施形態は、ディスクアレイスイッチのホストＩ／Ｆノードにおいて、フレームヘッダ４０１のみを変換し、フレームペイロード４０２は操作しない点、及び、ディスクアレイスイッチ、ホストＩ／Ｆ、ディスクアレイＩ／Ｆが二重化されていない点で第１実施形態と構成上相違する。したがって、各部の構成は、第１実施形態と大きく変わるところがなく、その詳細については説明を省略する。

【００７５】図１６において、各ディスクアレイサブセット１０は、複数の論理ユニット（ＬＵ）１１０で構成されている。各ＬＵ１１０は、独立ＬＵとして構成される。一般に、各ディスクアレイサブセット１０内のＬＵ１１０に割り当てられるＬＵＮは、０から始まる連続番号である。このため、ホスト３０に対して、ディスクアレイシステム1内のすべてのＬＵ１１０のＬＵＮを連続的に見せる場合には、第１実施形態と同様に、フレームペイロード４０２のＬＵＮフィールドを変換する必要がある。本実施形態では、各ディスクアレイサブセット１０のＬＵＮをそのままホスト３０に見せることで、フレームペイロード４０２の変換を不要とし、ディスクアレイスイッチの制御を簡単なものとしている。

【００７６】本実施形態のディスクアレイスイッチ２０は、ホストＩ／Ｆノード２０３ごとに特定のディスクアレイサブセット１０をアクセスできるものと仮定する。この場合、一つのホストＩ／Ｆ３１を使うと、１台のディスクアレイサブセット１０にあるＬＵ１１０のみがアクセス可能である。１台のホストから複数のディスクアレイサブセット１０のＬＵ１１０をアクセスしたい場合には、そのホストを複数のホストＩ／Ｆノード２０３に接続する。また、複数のホスト３０から１台のディスクアレイサブセット１０のＬＵ１１０をアクセスできるようにする場合は、同一のホストＩ／Ｆノード２０３にループトポロジーや、ファブリックトポロジー等を用い、複数のホスト３０を接続する。このように構成すると、１台のホスト３０から１つのＬＵ１１０をアクセスする際に、ホストＩ／Ｆノード２０３のD_ID毎にディスクアレイサブセット１０が確定することになるため、各ＬＵのＬＵＮをそのままホスト３０に見せることが可能である。

【００７７】本実施形態では、上述した理由により、ホスト３０に、各ディスクアレイサブセット１０内のＬＵ１１０のＬＵＮをそのままホスト３０に見せているため、ディスクアレイスイッチ２０におけるＬＵＮの変換は不要となる。このため、ディスクアレイスイッチ２０は、ホスト３０からフレームを受信すると、フレームヘッダ４０１のみを第１実施例と同様にして変換し、フレームペイロード４０２は変換せずにディスクアレイサブセット１０に転送する。本実施形態における各部の動作は、フレームペイロード４０２の変換が行われないことを除くと第１実施形態と同様であるので、ここでは詳細な説明を省略する。本実施形態によれば、ディスクアレイスイッチ２０の開発を容易にできる。

【００７８】［第３実施形態］第２実施形態では、ディスクアレイスイッチのホストＩ／Ｆノードにおいて、フレームヘッダのみを変換しているが、以下に説明する第３実施形態ではフレームヘッダも含め、フレームの変換を行わない形態について説明する。本実施形態のコンピュータシステムは、図１に示す第１実施形態におけるコンピュータシステムと同様に構成される。

【００７９】第１、および第２実施形態では、ホスト３０に対し、ディスクアレイサブセット１０の台数や、ＬＵ１１０の構成等、ディスクアレイシステム1の内部構成を隠蔽している。このため、ホスト３０からはディスクアレイシステム1が全体で１つの記憶装置として見える。これに対し、本実施形態では、ディスクアレイサブセット１０をそのままホスト３０に公開し、ホスト３０がフレームヘッダのD_IDとして直接ディスクアレイサブセットのポートのＩＤを使えるようにする。これにより、ディスクアレイスイッチは、フレームヘッダの情報に従ってフレームの転送を制御するだけで済み、従来技術におけるファイバチャネルのファブリック装置と同等のスイッチ装置をディスクアレイスイッチ２０に替えて利用することができる。

【００８０】ディスクアレイシステム構成管理手段７０は、ディスクアレイサブセット１０の通信コントローラ１０６、及びディスクアレイスイッチ２０の通信手段２０４と通信して各ディスクアレイサブセット１０及びディスクアレイスイッチ２０の構成情報を獲得し、あるいは、設定する。

【００８１】ディスクアレイスイッチ２０は、基本的には図３に示す第１実施形態におけるディスクアレイスイッチと同様の構成を有する。しかし、本実施形態では、ホスト３０が発行するフレームのフレームヘッダの情報をそのまま使ってフレームの転送を制御するため、第１

実施形態、あるいは第２実施形態でディスクアレイスイッチ２０のホストＩ／Ｆノード２０３、ディスクアレイＩ／Ｆノード２０２が有するＤＣＴ２０２７や、ＳＣ２０２２、ＳＰＣ２０２４等により実現されるフレームヘッダ等の変換の機能は不要となる。ディスクアレイスイッチ２０が有するクロスバスイッチ２０１は、フレームヘッダの情報に従ってホストＩ／Ｆノード２０３、及びディスクアレイＩ／Ｆノード２０２の間でファイバチャネルのフレームの転送を行う。

【００８２】本実施形態では、ディスクアレイシステムの構成をディスクアレイシステム構成管理手段７０で一括して管理するために、ディスクアレイ管理用テーブル（以下、このテーブルもＤＣＴと呼ぶ）をディスクアレイシステム構成管理手段７０に備える。ディスクアレイシステム構成管理手段７０が備えるＤＣＴは、図６、７に示す、システム構成テーブル20270とサブセット構成テーブル202720～202723の２つのテーブル群を含む。なお、本実施形態では、ホストＬＵは全てＩＬＵとして構成されるため、ホストＬＵ構成テーブル20271のLU Typeは全て「ＩＬＵ」となり、CLU Class、CLU Stripe Sizeは意味をなさない。

【００８３】管理者は、管理端末５を操作してディスクアレイシステム構成管理手段７０と通信し、ディスクアレイサブセット１０のディスク容量、ディスクユニットの台数等の情報を得て、ディスクアレイサブセット１０のＬＵ１１０の設定、ＲＡＩＤレベルの設定等を行う。次に管理者は、管理端末５によりディスクアレイシステム構成管理手段７０と通信し、ディスクアレイスイッチ２０を制御して、各ホスト３０とディスクアレイサブセット２０間の関係情報を設定する。

【００８４】以上の操作により、ディスクアレイシステム1の構成が確立し、ホスト３０から管理者が望む通りにＬＵ１１０が見えるようになる。ディスクアレイ構成管理手段７０は以上の設定情報を保存し、管理者からの操作に応じ構成の確認や、構成の変更を行うことができる。

【００８５】本実施形態によれば、ひとたびディスクアレイシステム1を構成すれば、管理者からディスクアレイスイッチ２０の存在を認識させることが無く、複数のディスクアレイサブシステムを１台のディスクアレイシステムと同様に扱うことができる。また、本実施形態によれば、ディスクアレイスイッチ２０とディスクアレイサブセット１０は、同一の操作環境によって統一的に操作することができ、その構成確認や、構成変更も容易になる。さらに、本実施形態によれば、従来使用していたディスクアレイシステムを本実施形態におけるディスクアレイシステムに置き換える場合に、ホスト３０の設定を変更することなく、ディスクアレイシステム1の構成をそれまで使用していたディスクアレイシステムの構成に合わせることができ、互換性を維持できる。

【００８６】［第４実施形態］以上説明した第１から第３の実施形態では、ホストＩ／Ｆにファイバチャネルを使用している。以下に説明する実施形態では、ファイバチャネル以外のインタフェースが混在した形態について説明する。

【００８７】図１７は、ホストＩ／ＦがパラレルＳＣＳＩである場合のホストＩ／Ｆノード２０３内部のＩＣ２０２３の一構成例を示す。20230はパラレルＳＣＳＩのプロトコル制御を行うＳＣＳＩプロトコルコントローラ（ＳＰＣ）、20233はファイバチャネルのプロトコル制御を行うファイバチャネルプロトコルコントローラ（ＦＰＣ）、20231はパラレルＳＣＳＩとファイバチャネルのシリアルＳＣＳＩをプロトコル変換するプロトコル変換プロセッサ（ＰＥＰ）、20232はプロトコル変換中データを一時保存するバッファ（ＢＵＦ）である。

【００８８】本実施形態において、ホスト３０は、ディスクアレイＩ／Ｆノード２０３に対してＳＣＳＩコマンドを発行する。リードコマンドの場合、ＳＰＣ20230は、これをＢＵＦ20232に格納し、ＰＥＰ20231に割り込みでコマンドの受信を報告する。ＰＥＰ20231は、ＢＵＦ20232に格納されたコマンドを利用し、ＦＰＣ20233へのコマンドに変換し、ＦＰＣ20233に送る。ＦＰＣ20233は、このコマンドを受信すると、フレーム形式に変換し、ＳＣ２０２２に引き渡す。この際、エクスチェンジＩＤ、シーケンスＩＤ、ソースＩＤ、デスティネイションＩＤは、以降の処理が可能なようにＰＥＰ20231により付加される。あとのコマンド処理は、第１実施形態と同様に行われる。

【００８９】ディスクアレイサブセット１０は、データの準備が完了すると、データ転送準備完了フレームの発行、データ転送、正常終了後ステータスフレームの発行を実施する。ディスクアレイサブセット１０からＩＣ２０２３までの間では、フレームヘッダ４０１やフレームペイロード４０２が必要に応じ変換されながら、各種フレームの転送が行われる。ＩＣ２０２３のＦＰＣ20233は、データ転送準備完了フレームを受信し、続いてデータを受信してＢＵＦ20232に格納し、続けて正常に転送が終わったならば、ステータスフレームを受信し、ＰＴＰ20231に割り込みをかけてデータの転送完了を報告する。ＰＴＰ20231は、割り込みを受けると、ＳＰＣ20230を起動し、ホスト３０に対しデータ転送を開始するよう指示する。ＳＰＣ20230はホスト３０にデータを送信し、正常終了を確認するとＰＴＰ20231に対し割り込みで正常終了を報告する。

【００９０】ここでは、ファイバチャネル以外のホストＩ／Ｆの例としてパラレルＳＣＳＩを示したが、他のインタフェース、例えば、メインフレームへのホストＩ／ＦであるESCON等に対しても同様に適用することが可能である。ディスクアレイスイッチ２０のホストＩ／Ｆノード２０３として、例えば、ファイバチャネル、パラレ

ルＳＣＳＩ、及びESCONに対応したホストＩ／Ｆノードを設けることで、１台のディスクアレイシステム1に、メインフレームと、パーソナルコンピュータ、ワークステーション等のいわゆるオープンシステムの両方を混在させて接続することが可能である。本実施形態では、ディスクアレイＩ／Ｆとしては、第１から第３実施形態と同様、ファイバチャネルを用いているが、ディスクアレイＩ／Ｆに対しても任意のＩ／Ｆを使用することが可能である。

【００９１】［第５実施形態］次に、ディスクアレイシステム1の構成管理の方法について、第５実施形態として説明する。図１８は、本実施形態のシステム構成図である。本実施形態では、ホスト３０が4台設けられている。ホスト“＃０”、“＃１”とディスクアレイシステム1の間のＩ／Ｆ３０はファイバチャネル、ホスト“＃２”とディスクアレイシステム1の間は、パラレルＳＣＳＩ（Ultra SCSI）、ホスト“＃３”とディスクアレイシステム1の間は、パラレルＳＣＳＩ（Ultra2 SCSI）で接続されている。

【００９２】パラレルＳＣＳＩのディスクアレイスイッチ２０への接続は第４実施形態と同様に行われる。ディスクアレイシステム1は、４台のディスクアレイサブセット３０を有する。ディスクアレイサブセット“＃０”には４つの独立ＬＵ、ディスクアレイサブセット“＃１”には２つの独立ＬＵがそれぞれ構成されている。ディスクアレイサブセット“＃２”と“＃３”で１つの統合ＬＵが構成されている。本実施形態では、第１実施形態と同様、ホスト３０に対しディスクアレイサブセット１０を隠蔽し、ファイバチャネルのフレームを変換するものとする。各ＬＵに割り当てられるＬＵＮは、ディスクアレイサブセット“＃０”のＬＵから順に、ＬＵＮ＝０、１、２、・・・６までの７つである。

【００９３】図１９は、管理端末５の表示画面上に表示される画面の一例である。図は、ホストＩ／Ｆ３１と各論理ユニット（ＬＵ）との対応を示した論理接続構成画面である。

【００９４】論理接続構成画面５０には、各ホストＩ／Ｆ３１に関する情報３１００、各ＬＵ１１０に関する情報11000、ディスクアレイサブセット１０とＬＵ１１０の関係等が表示される。ホストＩ／Ｆ３１に関する情報としては、Ｉ／Ｆ種類、Ｉ／Ｆ速度、ステータス等が含まれる。ＬＵ１１０に関する情報としては、格納サブセット番号、ＬＵＮ、容量、ＲＡＩＤレベル、ステータス、情報、等が表示される。管理者はこの画面を参照することで、容易にディスクアレイシステム１の構成を管理することができる。

【００９５】論理接続構成画面５０上で、ホストＩ／ＦとＬＵの間に引かれている線は、各ホストＩ／Ｆ３１を経由してアクセス可能なＬＵ１１０を示している。ホストＩ／Ｆから線の引かれていないＬＵ１１０に対して、そのホストＩ／Ｆに接続するホスト３０からはアクセスできない。ホスト３０によって、扱うデータ形式が異なり、また使用者も異なることから、セキュリティ維持上、適切なアクセス制限を設けることが不可欠である。そこで、システムを設定する管理者が、この画面を用いて、各ＬＵ１１０とホストＩ／Ｆとの間のアクセス許可をあたえるか否かによって、アクセス制限を実施する。図において、例えば、ＬＵ“＃０”は、ホストＩ／Ｆ“＃０”および“＃１”からアクセス可能であるが、ホストＩ／Ｆ“＃２”、“＃３”からはアクセスできない。ＬＵ“＃４”は、ホストＩ／Ｆ“＃２”からのみアクセス可能である。

【００９６】このようなアクセス制限を実現するためアクセス制限情報は、ディスクアレイシステム構成管理手段７０からディスクアレイスイッチ２０に対して送信される。ディスクアレイスイッチ２０に送られたアクセス制限情報は、各ホストＩ／Ｆノード２０３に配信され、各ホストＩ／Ｆノード２０３のＤＣＴ２０２７に登録される。ホストにより、アクセスが制限されたＬＵに対するＬＵ存在有無の検査コマンドが発行された場合、各ホストＩ／Ｆノード２０３は、ＤＣＴ２０２７の検査を行い、検査コマンドに対し応答しないか、あるいは、エラーを返すことで、そのＬＵは、ホストからは認識されなくなる。ＬＵ存在有無の検査コマンドとしては、ＳＣＳＩプロトコルの場合、Test Unit Readyコマンドや、Inquiryコマンドが一般に用いられる。この検査なしに、リード／ライトが実施されることはないため、容易にアクセスの制限をかけることが可能である。

【００９７】本実施形態ではホストＩ／Ｆ３１毎にアクセス制限をかけているが、これを拡張することで、ホスト３０毎にアクセス制限をかけることも容易に実現できる。また、ホストＩ／Ｆ３１、ホスト３０、あるいは、アドレス空間を特定して、リードのみ可、ライトのみ可、リード／ライトとも可、リード／ライトとも不可といった、コマンドの種別に応じたアクセス制限をかけることもできる。この場合、アクセス制限情報としてホストＩ／Ｆ番号、ホストＩＤ、アドレス空間、制限コマンド等を指定してディスクアレイスイッチ２０に制限を設定する。

【００９８】次に、新たなディスクアレイサブセット１０の追加について説明する。ディスクアレイサブセット１０を新規に追加する場合、管理者は、ディスクアレイスイッチ２０の空いているディスクアレイＩ／Ｆノード２０２に追加するディスクアレイサブセット１０を接続する。つづけて、管理者は、管理端末５を操作し、論理接続構成画面５０に表示されている「最新状態を反映」ボタン５００１を押下する。この操作に応答して、未設定のディスクアレイサブセットを表す絵が画面上に表示される（図示せず）。このディスクアレイサブセットの絵が選択されると、ディスクアレイサブセットの設

定画面が現れる。管理者は、表示された設定画面上で、新規に追加されたディスクアレイサブセットの各種設定を実施する。ここで設定される項目にはＬＵの構成、ＲＡＩＤレベル等がある。続けて、図１９の論理接続構成図の画面に切り替えると、新規ディスクアレイサブセットとＬＵが現れる。以降、ホストＩ／Ｆ３１毎に対するアクセス制限を設定し、「設定実行」ボタン５００２を押下すると、ディスクアレイスイッチ２０に対し、アクセス制限情報、およびディスクアレイサブセット、ＬＵの情報が転送され、設定が実行される。

【００９９】各ディスクアレイサブセット１０にＬＵ１１０を追加する際の手順も上述した手順で行われる。また、ディスクアレイサブセット、およびＬＵの削除についてもほぼ同様の手順で行われる。異なる点は、管理者が各削除部位を画面上で選択して「削除」ボタン５００３を押下し、適切な確認が行われたのち、実行される点である。以上のように、管理端末７０を用いることで、管理者はディスクアレイシステム全体を一元的に管理できる。

【０１００】［第６実施形態］次に、ディスクアレイスイッチ２０によるミラーリングの処理について、第６実施形態として説明する。ここで説明するミラーリングとは、２台のディスクアレイサブセットの２つの独立ＬＵにより二重書きをサポートする方法であり、ディスクアレイサブセットのコントローラまで含めた二重化である。従って、信頼性は、ディスクのみの二重化とは異なる。

【０１０１】本実施形態におけるシステムの構成は図１に示すものと同じである。図１に示す構成おいて、ディスクアレイサブセット“＃０”と“＃１”は全く同一のＬＵ構成を備えており、この２つのディスクアレイサブセットがホスト３０からは１つのディスクアレイとして見えるものとする。便宜上、ミラーリングされたディスクアレイサブセットのペアの番号を“＃０１”と呼ぶ。また、各ディスクアレイサブセットのＬＵ“＃０”とＬＵ“＃１”によってミラーリングペアが形成され、このＬＵのペアを便宜上、ＬＵ“＃０１”と呼ぶ。ＤＣＴ２０２７のホストＬＵ構成テーブル20271上でＬＵ＃０１を管理するための情報は、CLU Classに「Mirrored」が設定され、LU Info.として、ＬＵ＃０とＬＵ＃１に関する情報が設定される。その他の各部の構成は第１実施形態と同様である。

【０１０２】本実施形態における各部の動作は、第１実施例とほぼ同様である。以下、第１実施形態と相違する点について、ディスクアレイスイッチ２０のホストＩ／Ｆノード２０３の動作を中心に説明する。図２０は、本実施形態におけるライト動作時に転送されるフレームのシーケンスを示す模式図、図２１、２２は、ライト動作時におけるホストＩ／Ｆノード２０３による処理の流れを示すフローチャートである。

【０１０３】ライト動作時、ホスト３０が発行したライトコマンドフレーム（FCP_CMD）は、ＩＣ２０２３により受信される（図２０の矢印（ａ）：ステップ21001）。ＩＣ２０２３により受信されたライトコマンドフレームは、第１実施形態で説明したリード動作時におけるステップ20002　20005と同様に処理される（ステップ21002 - 21005）。

【０１０４】ＳＣ２０２２は、ＳＰ２０２１を使ってＤＣＴ２０２７を検索し、ミラー化されたディスクアレイサブセット“＃０１”のＬＵ“＃０１”へのライトアクセス要求であることを認識する（ステップ21006）。ＳＣ２０２２は、ＦＢ２０２５上に、受信したコマンドフレームの複製を作成する（ステップ21007）。ＳＣ２０２２は、ＤＣＴ２０２７に設定されている構成情報に基づいてコマンドフレームの変換を行い、ＬＵ“＃０”とＬＵ“＃１”の両者への別々のコマンドフレームを作成する（ステップ21008）。ここで、ＬＵ“＃０”を主ＬＵ、ＬＵ“＃１”を従ＬＵと呼び、コマンドフレームにもそれぞれ主コマンドフレーム、従コマンドフレームと呼ぶ。そして、両者別々にＥＴ２０２６にエクスチェンジ情報を格納し、ディスクアレイサブセット“＃０”およびディスクアレイサブセット“＃１”に対し作成したコマンドフレームを発行する（図２０の矢印（ｂ０）（ｂ１）：ステップ21009）。

【０１０５】各ディスクアレイサブセット“＃０”、“＃１”は、コマンドフレームを受信し、それぞれ独立にデータ転送準備完了フレーム（FCP_XFER_RDY）をディスクアレイスイッチ２０に送信する（図２０の矢印（ｃ０）（ｃ１））。ディスクアレイスイッチ２０では、ホストＩ／Ｆノード２０３が、第１実施形態におけるリード動作のステップ20011　20013と同様の処理により転送されてきたデータ転送準備完了フレームを処理する（ステップ21011 - 21013）。

【０１０６】各ディスクアレイサブセットからのデータ転送準備完了フレームがそろった段階で（ステップ21014）、ＳＣ２０２２は、主データ転送準備完了フレームに対する変換を実施し（ステップ21015）、ＩＣ２０２３により変換後のフレームをホスト３０に送信する（図２０の矢印（ｄ）：ステップ21015）。

【０１０７】ホスト３０は、データ転送準備完了フレームを受信した後、ライトデータ送信のため、データフレーム（FCP_DATA）をディスクアレイスイッチ２０に送信する（図２０の矢印（ｅ））。ホスト３０からのデータフレームは、ＩＣ２０２３により受信されると（ステップ21031）、リードコマンドフレームやライトコマンドフレームと同様に、ＦＢ２０２５に格納され、ＣＲＣ検査、フレームヘッダの解析が行われる（ステップ21032、21033）。フレームヘッダの解析結果に基づき、ＥＴ２０２６がＳＰ２０２１により検索され、エクスチェンジ情報が獲得される（ステップ21034）。

【０１０８】ＳＣ２０２２は、ライトコマンドフレームのときと同様に複製を作成し（ステップ21035）、その一方をディスクアレイサブセット“＃０”内のＬＵ“＃０”に、他方をディスクアレイサブセット“＃１”内のＬＵ“＃１”に向けて送信する（図２０の矢印（ｆ０）（ｆ１）：ステップ21037）。

【０１０９】ディスクアレイサブセット“＃０”、“＃１”は、各々、データフレームを受信し、ディスクユニット１０４に対しそれぞれライトし、ステータスフレーム（FCP_RSP）をディスクアレイスイッチ２０に送信する。

【０１１０】ＳＣ２０２２は、ディスクアレイサブセット“＃０”、“＃１”それぞれからステータスフレームを受信すると、それらのステータスフレームから拡張ヘッダを外してフレームヘッダを再現し、ＥＴ２０２６からエクスチェンジ情報を獲得する（ステップ21041、21042）。

【０１１１】ディスクアレイサブセット“＃０”、“＃１”の両者からのステータスフレームが揃うと（ステップ21043）、ステータスが正常終了であることを確認のうえ、ＬＵ“＃０”からの主ステータスフレームに対する変換を行い（ステップ21044）、従ステータスフレーム消去する（ステップ21045）。そして、ＩＣ２０２３は、正常終了を報告するためのコマンドフレームをホストに送信する（図２０の矢印（ｈ）：ステップ21046）。最後にＳＰ２０２１は、ＥＴ２０２６のエクスチェンジ情報を消去する（ステップ21047）。

【０１１２】以上でミラーリング構成におけるライト処理が終了する。ミラーリングされたＬＵ“＃０１”に対するリード処理は、データの転送方向が異なるだけで、上述したライト処理とほぼ同様に行われるが、ライトとは異なり、２台のディスクアレイサブセットにリードコマンドを発行する必要はなく、どちらか一方に対してコマンドフレームを発行すればよい。たとえば、常に主ＬＵに対してコマンドフレームを発行してもよいが、高速化のため、主／従双方のＬＵに対して、交互にコマンドフレームを発行するなどにより、負荷を分散すると有効である。

【０１１３】上述した処理では、ステップ21014、及びステップ21043で２台のディスクアレイサブセット“＃０”、“＃１”の応答を待ち、両者の同期をとって処理が進められる。このような制御では、双方のディスクアレイサブセットでの処理の成功が確認されてから処理が進むため、エラー発生時の対応が容易になる。その一方で、全体の処理速度が、どちらか遅いほうの応答に依存してしまうため、性能が低下するという欠点がある。

【０１１４】この問題を解決するため、ディスクアレイスイッチにおいて、ディスクアレイサブセットの応答を待たずに次の処理に進んだり、ディスクアレイサブセットのどちらか一方からの応答があった時点で次の処理に進む「非同期型」の制御をすることも可能である。非同期型の制御を行った場合のフレームシーケンスの一例を、図２０において破線矢印で示す。

【０１１５】破線矢印で示されるフレームシーケンスでは、ステップ21016で行われるホストへのデータ転送準備完了フレームの送信が、ステップ21009の処理の後、ディスクアレイサブセット１０からのデータ転送準備完了フレームを待たずに実施される。この場合、ホストに送信されるデータ転送準備完了フレームは、ディスクアレイスイッチ２０のＳＣ２０２２により生成される（破線矢印（ｄ’））。

【０１１６】ホスト３０からは、破線矢印（ｅ’）で示されるタイミングでデータフレームがディスクアレイスイッチ２０に転送される。ディスクアレイスイッチ２０では、このデータフレームが一旦ＦＢ２０２５に格納される。ＳＣ２０２２は、ディスクアレイサブセット１０からのデータ転送準備完了フレームの受信に応答して、データ転送準備完了フレームが送られてきたディスクアレイサブセット１０に対し、ＦＢ２０２５に保持されたデータフレームを転送する（破線矢印（ｆ０’）、（ｆ１’））。

【０１１７】ディスクアレイスイッチ２０からホスト３０への終了報告は、双方のディスクアレイサブシステム１０からの報告（破線矢印（ｇ０’）、（ｇ０’））があった時点でおこなわれる（破線矢印（ｈ’））。このような処理により、図２０に示される時間Ｔａの分だけ処理時間を短縮することが可能である。

【０１１８】ディスクアレイスイッチ２０とディスクアレイサブセット１０間のフレーム転送の途中でエラーが発生した場合、以下の処理が実施される。

【０１１９】実行中の処理がライト処理の場合、エラーが発生したＬＵに対し、リトライ処理が行われる。リトライが成功すれば、処理はそのまま継続される。あらかじめ設定された規定の回数のリトライが失敗した場合、ディスクアレイスイッチ２０は、このディスクアレイサブセット１０（もしくはＬＵ）に対するアクセスを禁止し、そのことを示す情報をＤＣＴ２０２７に登録する。また、ディスクアレイスイッチ２０は、ＭＰ２００、通信コントローラ２０４を経由して、ディスクシステム構成手段７０にそのことを通知する。

【０１２０】ディスクシステム構成手段７０は、この通知に応答して管理端末５にアラームを発行する。これにより管理者は、トラブルが発生したことを認識できる。その後、ディスクアレイスイッチ２０は、正常なディスクアレイサブセットを用いて運転を継続する。ホスト３０は、エラーが発生したことを認識することはなく、処理を継続できる。

【０１２１】本実施形態によれば、２台のディスクアレイサブシステムでミラー構成を実現できるので、ディスクの耐障害性を上げることことができる。また、ディス

クアレイコントローラ、ディスクアレイＩ／Ｆ、及びディスクアレイＩ／Ｆノードの耐障害性を上げることができ、内部バスの二重化等するくとなくディスクアレイシステム全体の信頼性を向上させることができる。

【０１２２】［第７実施形態］次に、３台以上のディスクアレイサブセット１０を統合し、１台の論理的なディスクアレイサブセットのグループを構成する方法について説明する。本実施形態では、複数のディスクアレイサブセット１０にデータを分散して格納する。これにより、ディスクアレイサブセットへのアクセスを分散させ、特定のディスクアレイサブセットへのアクセスの集中を抑止することで、トータルスループットを向上させる。本実施形態では、ディスクアレイスイッチによりこのようなストライピング処理を実施する。

【０１２３】図２３は、本実施形態におけるディスクアレイシステム1のアドレスマップである。ディスクアレイサブセット１０のアドレス空間は、ストレイプサイズＳでストライピングされている。ホストから見たディスクアレイシステム１のアドレス空間は、ストライプサイズＳ毎に、ディスクアレイサブセット“＃０”、“＃１”、“＃２”、“＃３”に分散されている。ストライプサイズＳのサイズは任意であるが、あまり小さくない方がよい。ストライプサイズＳが小さすぎると、アクセスすべきデータが複数のストライプに属するストライプまたぎが発生したときに、その処理にオーバヘッドが発生するおそれがある。ストライプサイズＳを大きくすると、ストライプまたぎが発生する確率が減少するので性能向上のためには好ましい。ＬＵの数は任意に設定することができる。

【０１２４】以下、本実施形態におけるホストＩ／Ｆノード２０３の動作について、図２４に示す動作フローチャートを参照しつつ第１実施形態との相違点に着目して説明する説明する。なお、本実施形態では、ＤＣＴ２０２７のホストＬＵ構成テーブル20271上で、ストライピングされたホストＬＵに関する情報のCLU Classには「Striped」が、CLU Stripe Sizeにはストライプサイズ「Ｓ」が設定される。

【０１２５】ホスト３０がコマンドフレームを発行すると、ディスクアレイスイッチ２０は、ホストＩ／Ｆノード２０３のＩＣ２０２３でこれを受信する（ステップ２２００１）、ＳＣ２０２２は、ＩＣ２０２３からこのコマンドフレームを受け取り、ＳＰ２０２１を使ってＤＣＴ２０２７を検索し、ストライピングする必要があることを認識する（ステップ２２００５）。

【０１２６】次に、ＳＣ２０２２は、ＳＰ２０２１によりＤＣＴ２０２７を検索し、ストライプサイズＳを含む構成情報から、アクセスの対象となるデータが属するストライプのストライプ番号を求め、このストライプがどのディスクアレイサブセット１０に格納されているか特定する（ステップ22006）。この際、ストライプまたぎが発生する可能性があるが、この場合の処理については後述する。ストライプまたぎが発生しない場合、ＳＰ２０２１の計算結果に基づき、ＳＣ２０２２はコマンドフレームに対し変換を施し（ステップ22007）、エクスチェンジ情報をＥＴ２０２６に格納する（ステップ22008）。以降は、第１実施形態と同様の処理が行われる。

【０１２７】ストライプまたぎが発生した場合、ＳＰ２０２１は、２つのコマンドフレームを生成する。この生成は、例えば、ホスト３０が発行したコマンドフレームを複製することで行われる。生成するコマンドフレームのフレームヘッダ、フレームペイロード等は、新規に設定する。第６実施形態と同様、ＳＣ２０２２でコマンドフレームの複製を作成した後、変換を実施することも可能であるが、ここでは、ＳＰ２０２１により新規に作成されるものとする。ＳＣ２０２２は、２つのコマンドフレームが生成されると、これらを各ディスクアレイサブセット１０に送信する。

【０１２８】この後、第１実施形態と同様にデータ転送が実施される。ここで、本実施形態では、第１実施形態、あるいは第６実施形態と異なり、データ自体を１台のホスト３０と２台のディスクアレイサブセット１０間で転送する必要がある。たとえば、リード処理の場合、２台のディスクアレイサブセット１０から転送されるデータフレームは、すべてホスト３０に転送する必要がある。この際ＳＣ２０２２は、各ディスクアレイサブセット１０から転送されてくるデータフレームに対し、ＥＴ２０２６に登録されたエクスチェンジ情報に従い、適切な順番で、適切なエクスチェンジ情報を付加してホスト３０に送信する。

【０１２９】ライト処理の場合は、コマンドフレームの場合と同様、２つのデータフレームに分割して、該当するディスクアレイサブセット１０に転送する。なお、データフレームの順序制御は、ホスト、あるいはディスクアレイサブセットがアウトオブオーダー（Out of Order）機能と呼ばれる、順不同処理に対応しているならば必須ではない。

【０１３０】最後に、すべてのデータ転送が完了し、ディスクアレイスイッチ２０が２つのステータスフレームをディスクアレイサブセット１０から受信すると、ＳＰ２０２１（あるいはＳＣ２０２２）は、ホスト３０へのステータスフレームを作成し、これをＩＣ２０２３によりホスト３０に送信する。

【０１３１】本実施形態によれば、アクセスを複数のディスクアレイサブセットに分散することができるので、トータルとしてスループットを向上させることができるとともに、アクセスレイテンシも平均的に低減させることが可能である。

【０１３２】［第８実施形態］次に、２台のディスクアレイシステム（またはディスクアレイサブセット）間における複製の作成について、第８実施形態として説明す

る。ここで説明するようなシステムは、２台のディスクアレイシステムの一方を遠隔地に配置し、天災等による他方のディスクアレイシステムの障害に対する耐性を備える。このような災害に対する対策をディザスタリカバリと呼び、遠隔地のディスクアレイシステムとの間で行われる複製の作成のことをリモートコピーと呼ぶ。

【０１３３】第６実施形態で説明したミラーリングでは、地理的にほぼ同一の場所に設置されたディスクアレイサブセット１０でミラーを構成するので、ディスクアレイＩ／Ｆ２１はファイバチャネルでよい。しかし、リモートコピーを行うディスクアレイ（ディスクアレイサブセット）が１０ｋｍを越える遠隔地に設置される場合、中継なしでファイバチャネルによりフレームを転送する事ができない。ディザスタリカバリに用いられる場合、お互いの間の距離は通常数百ｋｍ以上となる、このため、ファイバチャネルでディスクアレイ間を接続することは実用上不可能であり、ＡＴＭ（Asynchronous Transfer Mode）等による高速公衆回線や衛星通信等が用いられる。

【０１３４】図２５は、本実施形態におけるディザスタリカバリシステムの構成例である。

【０１３５】８１はサイトＡ、８２はサイトＢであり、両サイトは、地理的な遠隔地に設置される。9は公衆回線であり、ＡＴＭパケットがここを通過する。サイトＡ８１、およびサイトＢ８２は、それぞれディスクアレイシステム1を有する。ここでは、サイトＡ８１が通常使用される常用サイトであり、サイトＢ８２はサイトＡ８１が災害等でダウンしたときに使用されるリモートディザスタリカバリサイトである。

【０１３６】サイトＡ８１のディスクアレイシステム１０のディスクアレイサブセット“＃０”、“＃１”の内容は、サイトＢ８２のディスクアレイシステム１０のリモートコピー用ディスクアレイサブセット“＃０”、“＃１”にコピーされる。ディスクアレイスイッチ２０のＩ／Ｆノードのうち、リモートサイトに接続するものはＡＴＭを用いて公衆回線9に接続されている。このノードをＡＴＭノード２０５と呼ぶ。ＡＴＭノード２０５は、図５に示すホストＩ／Ｆノードと同様に構成され、ＩＣ２０２３がＡＴＭ－ファイバチャネルの変換を行う。この変換は、第４実施形態におけるＳＣＳＩ－ファイバチャネルの変換と同様の方法により実現される。

【０１３７】本実施形態におけるリモートコピーの処理は、第６実施形態におけるミラーリングの処理と類似する。以下、第６実施形態におけるミラーリングの処理と異なる点について説明する。

【０１３８】ホスト３０がライトコマンドフレームを発行すると、サイトＡ８１のディスクアレイシステム１０は、第６実施形態における場合と同様にフレームの二重化を実施し、その一方を自身のディスクアレイサブセット10に転送する。他方のフレームは、ＡＴＭノード２０５によりファイバチャネルフレームからＡＴＭパケットに変換され、公衆回線9を介してサイトＢ８２に送られる。

【０１３９】サイトＢ８２では、ディスクアレイスイッチ２０のＡＴＭノード２０５がこのパケットを受信する。ＡＴＭノード２０５のＩＣ２０２３は、ＡＴＭパケットからファイバチャネルフレームを再現し、ＳＣ２０２２に転送する。ＳＣ２０２２は、ホスト３０からライトコマンドを受信したときと同様にフレーム変換を施し、リモートコピー用のディスクアレイサブセットに転送する。以降、データ転送準備完了フレーム、データフレーム、ステータスフレームのすべてにおいて、ＡＴＭノード２０５においてファイバチャネル－ＡＴＭ変換を行い、同様のフレーム転送処理を実施することにより、リモートコピーが実現できる。

【０１４０】ホスト３０がリードコマンドフレームを発行した際には、ディスクアレイスイッチ２０は、自サイトのディスクアレイサブセット１０に対してのみコマンドフレームを転送し、自サイトのディスクアレイサブセット１０からのみデータをリードする。このときの動作は、第１実施形態と同一となる。

【０１４１】本実施形態によれば、ユーザデータをリアルタイムでバックアップし、天災等によるサイト障害、ディスクアレイシステム障害に対する耐性を備えることができる。

【０１４２】［第９実施形態］次に、一台のディスクアレイサブセット１０に包含される複数のＬＵの統合について説明する。例えば、メインフレーム用のディスク装置は、過去のシステムとの互換性を維持するために、論理ボリュームのサイズの最大値が２ＧＢに設定されている。このようなディスクアレイシステムをオープンシステムでも共用する場合、ＬＵは論理ボリュームサイズの制限をそのまま受けることになり、小サイズのＬＵが多数ホストから見えることになる。このような方法では、大容量化が進展した場合に運用が困難になるという問題が生じる。そこで、ディスクアレイスイッチ２０の機能により、この論理ボリューム（すなわちＬＵ）を統合して一つの大きな統合ＬＵを構成することを考える。本実施形態では、統合ＬＵの作成をディスクアレイスイッチ２０で実施する。

【０１４３】本実施形態におけるＬＵの統合は、第１実施形態における複数のディスクアレイサブセット１０による統合ＬＵの作成と同一である。相違点は、同一のディスクアレイサブセット１０内の複数ＬＵによる統合であることだけである。ディスクアレイシステムとしての動作は、第１実施形態と全く同一となる。

【０１４４】このように、同一のディスクアレイサブセット１０に包含される複数のＬＵを統合して一つの大きなＬＵを作成することで、ホストから多数のＬＵを管理する必要がなくなり、運用性に優れ、管理コストを低減

したディスクアレイシステムを構築できる。

【０１４５】［第１０実施形態］次に、ディスクアレイスイッチ１０による交代パスの設定方法について、図２６を参照しつつ説明する。

【０１４６】図２６に示された計算機システムにおける各部の構成は、第１の実施形態と同様である。ここでは、２台のホスト３０が、各々異なるディスクアレイＩ／Ｆ２１を用いてディスクアレイサブセット１０をアクセスするとように構成していると仮定する。図では、ディスクアレイサブセット、ディスクアレイスイッチ２０のホストＩ／Ｆノード２０３およびディスクアレイＩ／Ｆノード２０２は、ここでの説明に必要な数しか示されていない。

【０１４７】ディスクアレイサブセット１０は、図２と同様の構成を有し、２つのディスクアレイＩ／Ｆコントローラはそれぞれ１台のディスクアレイスイッチ２０に接続している。ディスクアレイスイッチ２０の各ノードのＤＣＴ２２７には、ディスクアレイＩ／Ｆ２１の交代パスが設定される。交代パスとは、ある一つのパスに障害が発生した場合にもアクセス可能になるように設けられる代替のパスのことである。ここでは、ディスクアレイＩ／Ｆ“＃０”の交替パスをディスクアレイＩ／Ｆ“＃１”、ディスクアレイＩ／Ｆ“＃１”の交替パスをディスクアレイＩ／Ｆ“＃０”として設定しておく。同様に、ディスクアレイサブセット１０内の上位アダプタ間、キャッシュ・交代メモリ間、下位アダプタ間のそれぞれについても交代パスを設定しておく。

【０１４８】次に、図２６に示すように、ディスクアレイサブセット１の上位アダプタ“＃１”に接続するディスクアレイＩ／Ｆ２１が断線し、障害が発生したと仮定して、交替パスの設定動作を説明する。このとき、障害が発生したディスクアレイＩ／Ｆ２１を利用しているホスト“＃１”は、ディスクアレイサブセット１０にアクセスできなくなる。ディスクアレイスイッチ２０は、ディスクアレイサブセット１０との間のフレーム転送の異常を検出し、リトライ処理を実施しても回復しない場合、このパスに障害が発生したと認識する。

【０１４９】パスの障害が発生すると、ＳＰ２０２１は、ＤＣＴ２０２７にディスクアレイＩ／Ｆ“＃１”に障害が発生したことを登録し、交代パスとしてディスクアレイＩ／Ｆ“＃０”を使用することを登録する。以降、ホストＩ／Ｆノード２０３のＳＣ２０２２は、ホスト“＃１”からのフレームをディスクアレイＩ／Ｆ“＃０”に接続するディスクアレイＩ／Ｆノード２０２に転送するように動作する。

【０１５０】ディスクアレイサブセット１０の上位アダプタ１０１は、ホスト“＃１”からのコマンドを引き継いで処理する。また、ディスクアレイスイッチ２０は、ディスクアレイシステム構成管理手段７０に障害の発生を通知し、ディスクアレイシステム構成管理手段７０により管理者に障害の発生が通報される。

【０１５１】本実施形態によれば、パスに障害が発生した際の交替パスへの切り替えを、ホスト側に認識させることなく行うことができ、ホスト側の交代処理設定を不要にできる。これにより、システムの可用性を向上させることができる。

【０１５２】以上説明した各実施形態では、記憶メディアとして、すべてディスク装置を用いたディスクアレイシステムについて説明した。しかし、本発明は、これに限定されるものではなく、記憶メディアとしてディスク装置に限らず、光ディスク装置、テープ装置、ＤＶＤ装置、半導体記憶装置等を用いた場合にも同様に適用できる。

【０１５３】

【発明の効果】本発明によれば、計算機システムの規模、要求などに応じた記憶装置システムの拡張、信頼性の向上などを容易に実現することのできる記憶装置システムを実現することができる。

【図面の簡単な説明】

【図１】第１実施形態のコンピュータシステムの構成図である。

【図２】第１実施形態のディスクアレイサブセットの構成図である。

【図３】第１実施形態のディスクアレイスイッチの構成図である。

【図４】第１実施形態におけるディスクアレイスイッチのクロスバスイッチの構成図である。

【図５】第１実施形態におけるディスクアレイスイッチのホストＩ／Ｆノードの構成図である。

【図６】システム構成テーブルの構成図である。

【図７】サブセット構成テーブルの構成図である。

【図８】ファイバチャネルのフレームの構成図である。

【図９】ファイバチャネルのフレームヘッダの構成図である。

【図１０】ファイバチャネルのフレームペイロードの構成図である。

【図１１】ホストからのリード動作時にファイバチャネルを通して転送されるフレームのシーケンスを示す模式図である。

【図１２】ホストＬＵ、各ディスクアレイサブセットのＬＵ、及び各ディスクユニットの対応関係を示す模式図である。

【図１３】ライト処理時のホストＩ／Ｆノードにおける処理のフローチャートである。

【図１４】スイッチングパケットの構成図である。

【図１５】複数のディスクアレイスイッチをクラスタ接続したディスクアレイシステムの構成図である。

【図１６】第２実施形態におけるコンピュータシステムの構成図である。

【図１７】第４実施形態におけるディスクアレイスイッ

チのインタフェースコントローラの構成図である。

【図１８】第５実施形態におけるコンピュータシステムの構成図である。

【図１９】論理接続構成画面の表示例を示す画面構成図である。

【図２０】第６実施形態におけるフレームシーケンスを示す模式図である。

【図２１】第６実施形態のミラーリングライト処理時のホストＩ／Ｆノードにおける処理のフローチャートである。

【図２２】第６実施形態のミラーリングライト処理時のホストＩ／Ｆノードにおける処理のフローチャートである。

【図２３】第７実施形態におけるホストＬＵと各ディスクアレイサブセットのＬＵとの対応関係を示す模式図である。

【図２４】第７実施形態におけるホストＩ／Ｆノードの処理を示すフローチャートである。

【図２５】第８実施形態におけるディザスタリカバリシステムの構成図である。

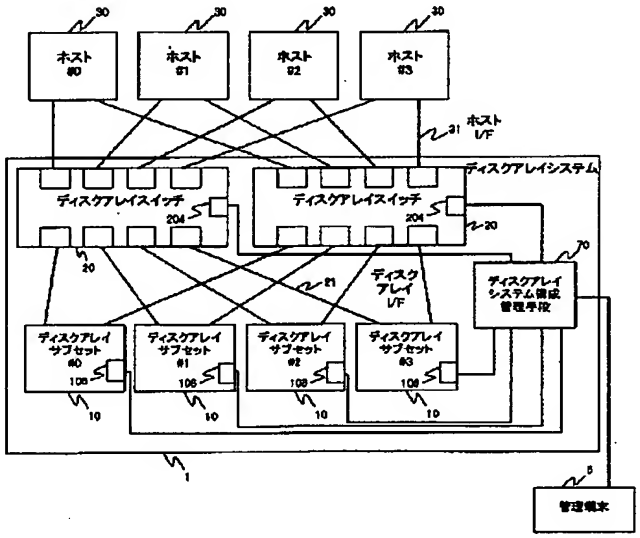
【図２６】交替パスの設定についての説明図である。

【符号の説明】

１…ディスクアレイシステム、５…管理端末、１０…ディスクアレイサブセット、２０…ディスクアレイスイッチ、３０…ホストコンピュータ、７０…ディスクアレイシステム構成管理手段、２００…管理プロセッサ、２０１…クロスバスイッチ、２０２…ディスクアレイＩ／Ｆノード、２０３…ホストＩ／Ｆノード、２０４…通信コントローラ。

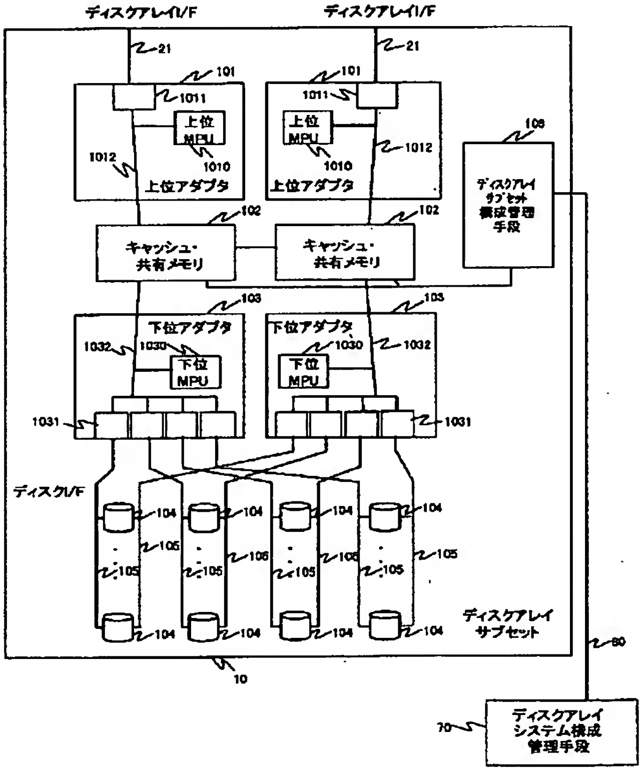
【図１】

図1



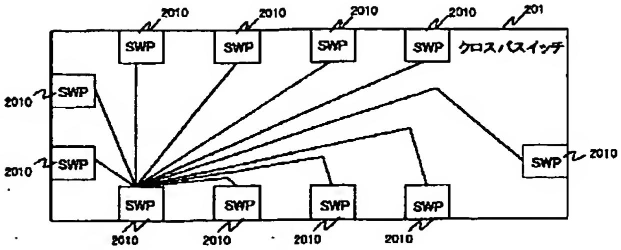
【図２】

図2



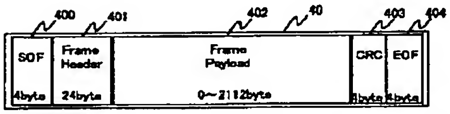
【図４】

図4



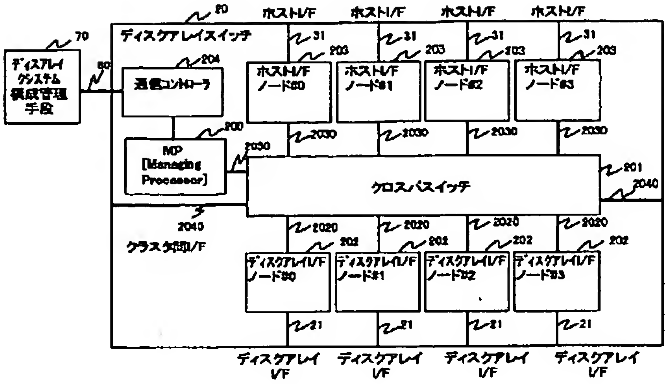
【図８】

図8



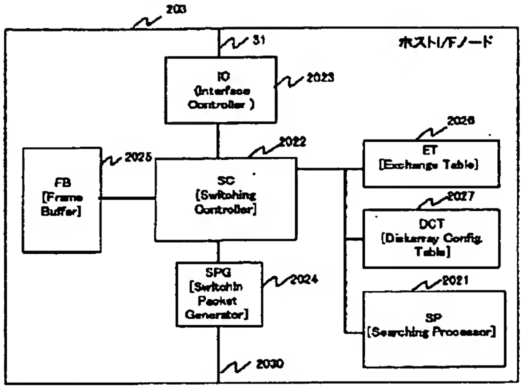
【図３】

図3



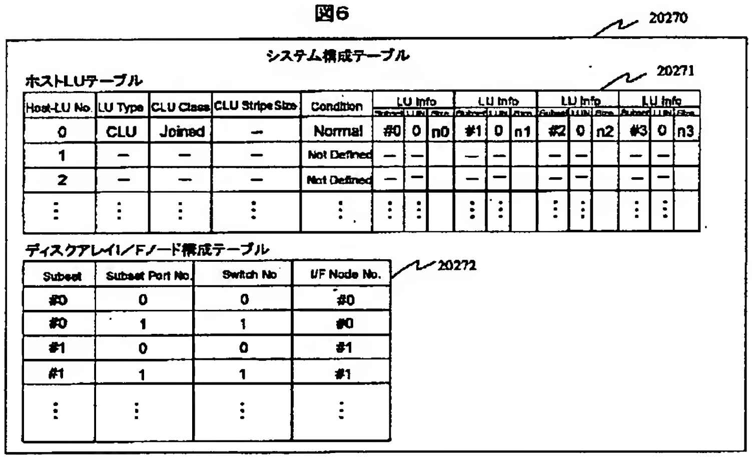
【図５】

図5



【図６】

図6


20270
20271
20272

システム構成テーブル

ホストLUテーブル

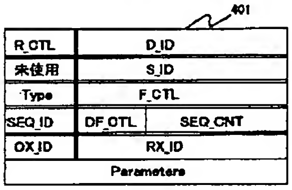
Host-LU No.	LU Type	CLU Class	CLU Stripe Size	Condition	LU Info			LU Info			LU Info			LU Info		
					Subset	LUN	Size	Subset	LUN	Size	Subset	LUN	Size	Subset	LUN	Size
0	CLU	Joined	−	Normal	#0	0	n0	#1	0	n1	#2	0	n2	#3	0	n3
1	−	−	−	Not Defined	−	−		−	−		−	−		−	−	
2	−	−	−	Not Defined	−	−		−	−		−	−		−	−	
⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮		⋮	⋮		⋮	⋮	

ディスクアレイI/Fノード構成テーブル

Subset	Subset Port No.	Switch No	I/F Node No.
#0	0	0	#0
#0	1	1	#0
#1	0	0	#1
#1	1	1	#1
⋮	⋮	⋮	⋮

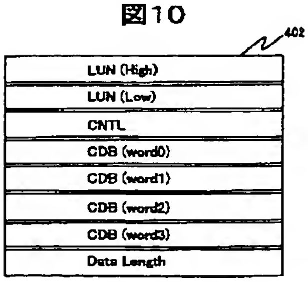
【図９】

図9



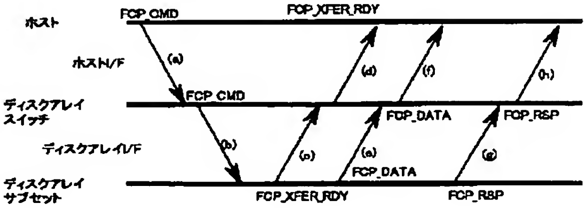
【図１０】

図10



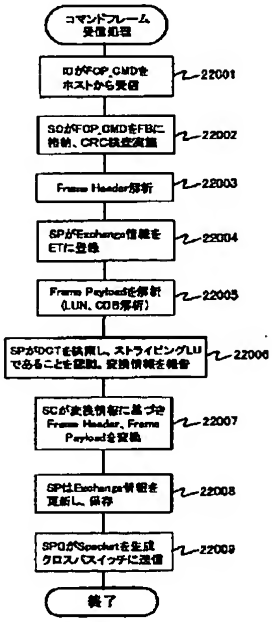
【図１１】

図11



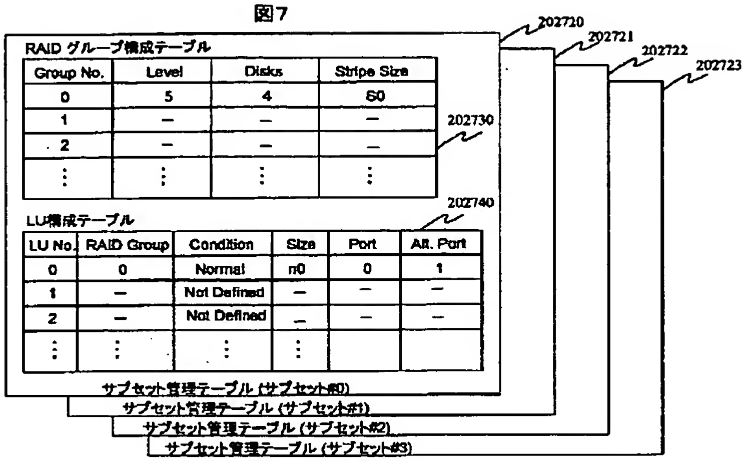
【図２４】

図24


コマンドフレーム受信処理
IOがFCP_CMDをホストから受信
22001
SCがFCP_CMDをFBに格納、CRC検査実施
22002
Frame Header解析
22003
SPがExchange情報をETに登録
22004
Frame Payloadを解析(LUN、CDB解析)
22005
SPがDCTを検索し、ストライピングLUであることを認識。変換情報を報告
22006
SCが変換情報に基づきFrame Header、Frame Payloadを変換
22007
SPはExchange情報を更新し、保存
22008
SPGがSpacketを生成クロスパスイッチに送信
22009
終了

【図７】

図7



RAID グループ構成テーブル

Group No.	Level	Disks	Stripe Size
0	5	4	50
1	—	—	—
2	—	—	—
⋮	⋮	⋮	⋮

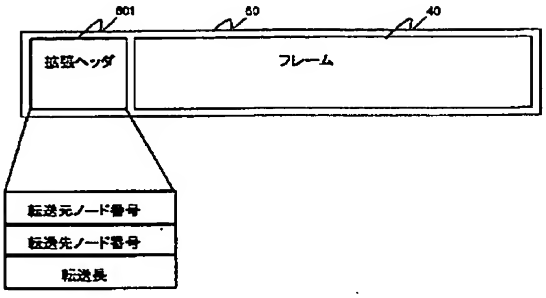
LU構成テーブル

LU No.	RAID Group	Condition	Size	Port	Alt. Port
0	0	Normal	n0	0	1
1	—	Not Defined	—	—	—
2	—	Not Defined	_	—	—
⋮	⋮	⋮	⋮		

サブセット管理テーブル (サブセット#0)
サブセット管理テーブル (サブセット#1)
サブセット管理テーブル (サブセット#2)
サブセット管理テーブル (サブセット#3)

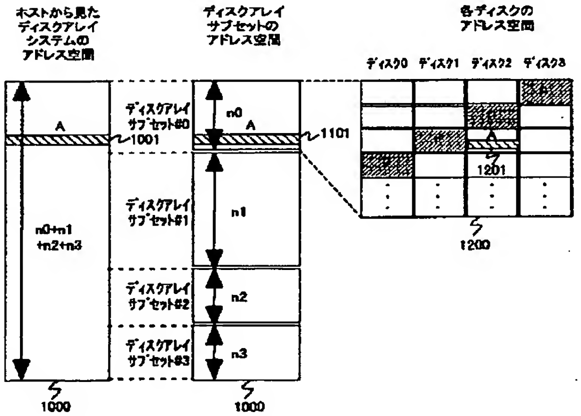
【図１４】

図14



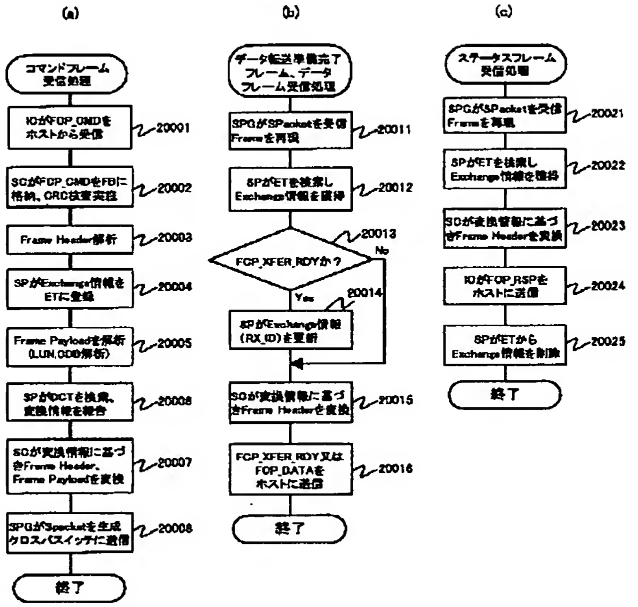
【図１２】

図12



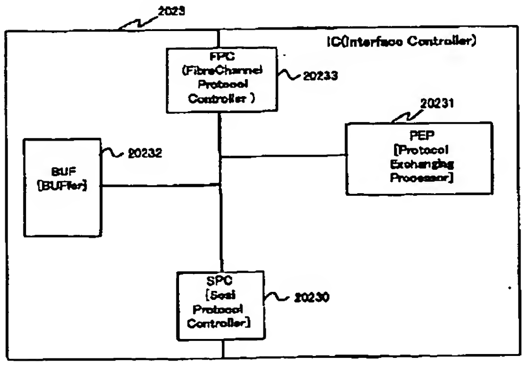
【図１３】

図13



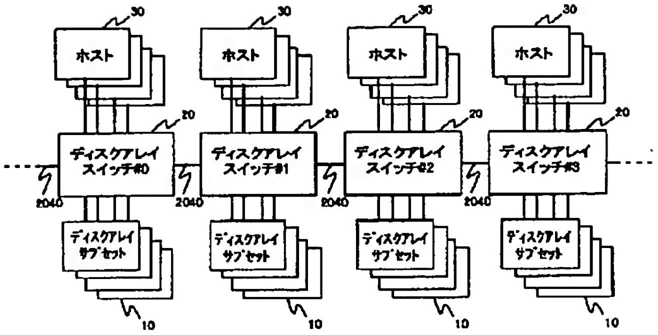
【図１７】

図17



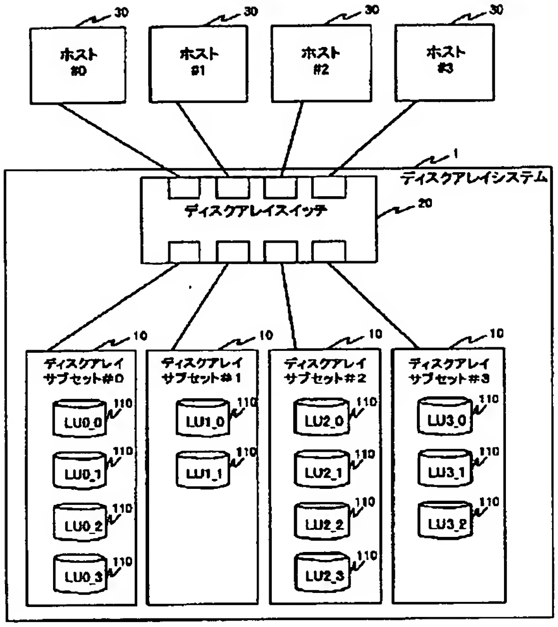
【図１５】

図15


30
ホスト
ホスト
ホスト
ホスト
20
ディスクアレイ スイッチ#0
ディスクアレイ スイッチ#1
ディスクアレイ スイッチ#2
ディスクアレイ スイッチ#3
2040
ディスクアレイ サブセット
10

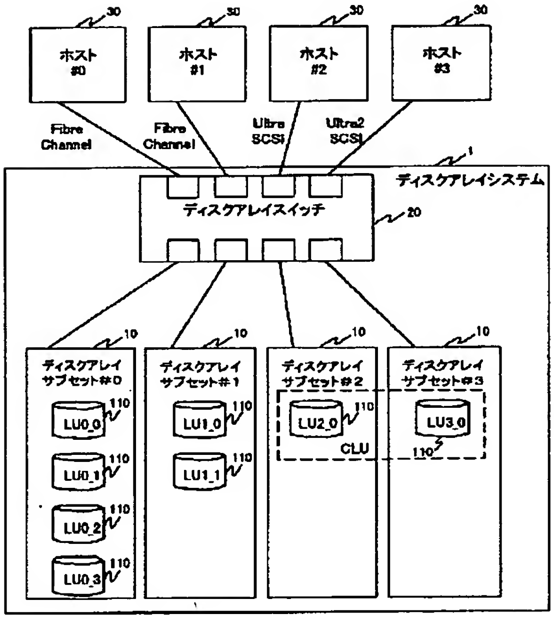
【図１６】

図16



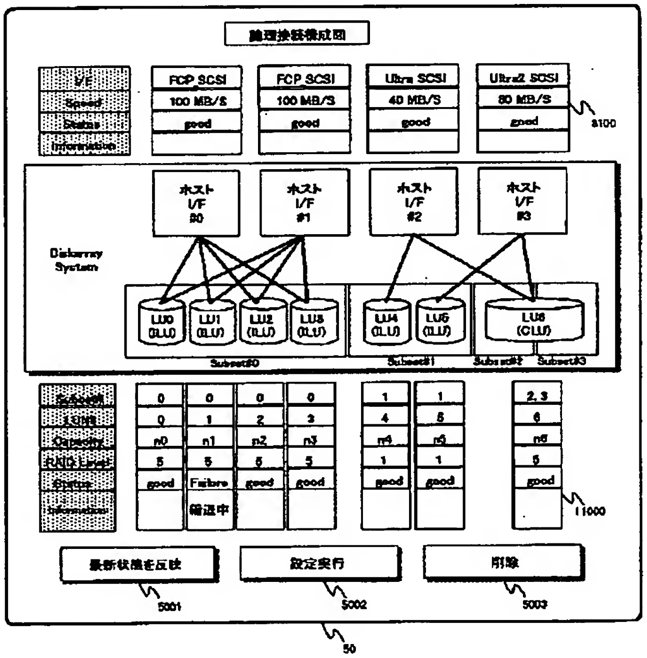
【図１８】

図18



【図１９】

図19

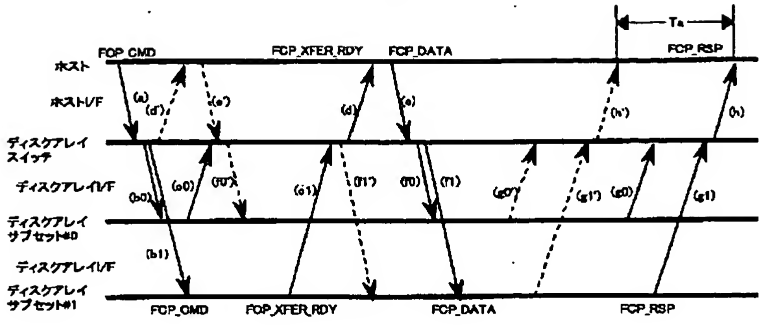


I/F	FCP SCSI	FCP SCSI	Ultra SCSI	Ultra2 SCSI
Speed	100 MB/S	100 MB/S	40 MB/S	80 MB/S
Status	good	good	good	good
Information				

Subset#	0	0	0	0	1	1	2, 3
LU#	0	1	2	3	4	5	6
Capacity	n0	n1	n2	n3	n4	n5	n6
RAID Level	5	5	5	5	1	1	5
Status	good	Failure	good	good	good	good	good
Information		縮退中					

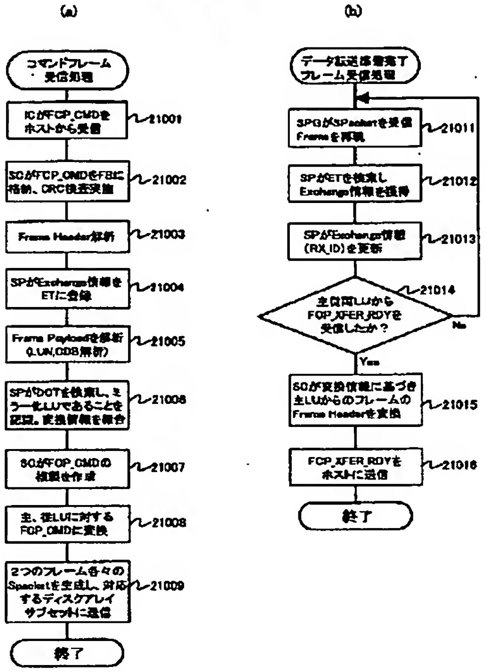
【図２０】

図20



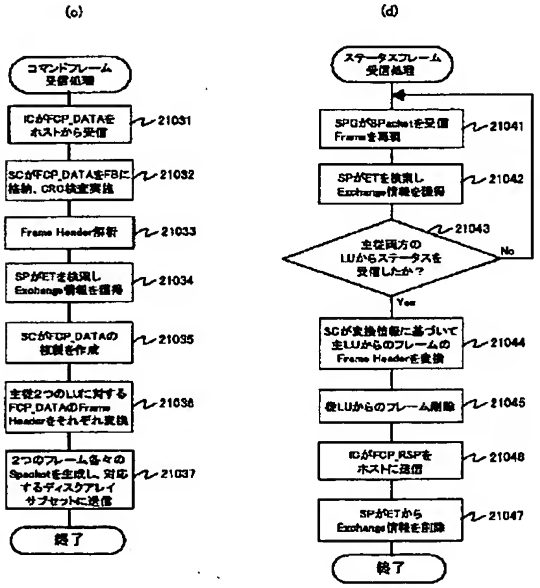
【図２１】

図21


(a)
コマンドフレーム受信処理
ICがFCP_CMDをホストから受信
21001
SCがFCP_CMDをFBに格納、CRC検査実施
21002
Frame Header解析
21003
SPがExchange情報をETに登録
21004
Frame Payloadを解析(LUN,CDB解析)
21005
SPがDOTを検索し、ミラー化LUであることを認識。変換情報を報告
21006
SCがFCP_CMDの複製を作成
21007
主、従LUに対するFCP_CMDに変換
21008
2つのフレーム各々のSpacketを生成し、対応するディスクアレイサブセットに送信
21009
終了
(b)
データ転送準備完了フレーム受信処理
SPGがSPacketを受信Frameを再現
21011
SPがETを検索しExchange情報を獲得
21012
SPがExchange情報(RX_ID)を更新
21013
主従両LUからFCP_XFER_RDYを受信したか？
21014
No
Yes
SCが変換情報に基づき主LUからのフレームのFrame Headerを変換
21015
FCP_XFER_RDYをホストに送信
21016
終了

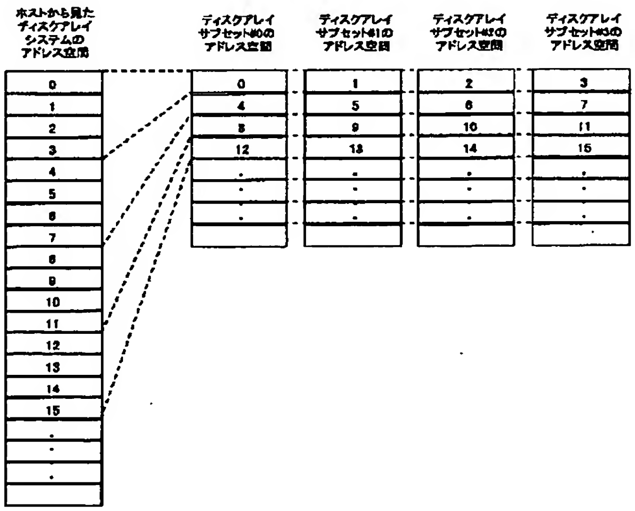
【図２２】

図22


(c)
コマンドフレーム受信処理
ICがFCP_DATAをホストから受信
21031
SCがFCP_DATAをFBに格納、CRC検査実施
21032
Frame Header解析
21033
SPがETを検索しExchange情報を獲得
21034
SCがFCP_DATAの複製を作成
21035
主従2つのLUに対するFCP_DATAのFrame Headerをそれぞれ変換
21036
2つのフレーム各々のSpacketを生成し、対応するディスクアレイサブセットに送信
21037
終了
(d)
ステータスフレーム受信処理
SPGがSPacketを受信Frameを再現
21041
SPがETを検索しExchange情報を獲得
21042
主従両方のLUからステータスを受信したか？
21043
No
Yes
SCが変換情報に基づいて主LUからのフレームのFrame Headerを変換
21044
従LUからのフレーム削除
21045
ICがFCP_RSPをホストに送信
21046
SPがETからExchange情報を削除
21047
終了

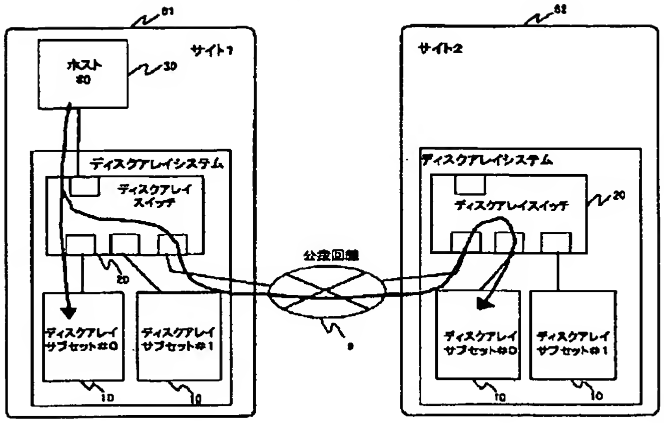
【図２３】

図23


ホストから見たディスクアレイシステムのアドレス空間
ディスクアレイサブセット#0のアドレス空間
ディスクアレイサブセット#1のアドレス空間
ディスクアレイサブセット#2のアドレス空間
ディスクアレイサブセット#3のアドレス空間
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
0
4
8
12
1
5
9
13
2
6
10
14
3
7
11
15

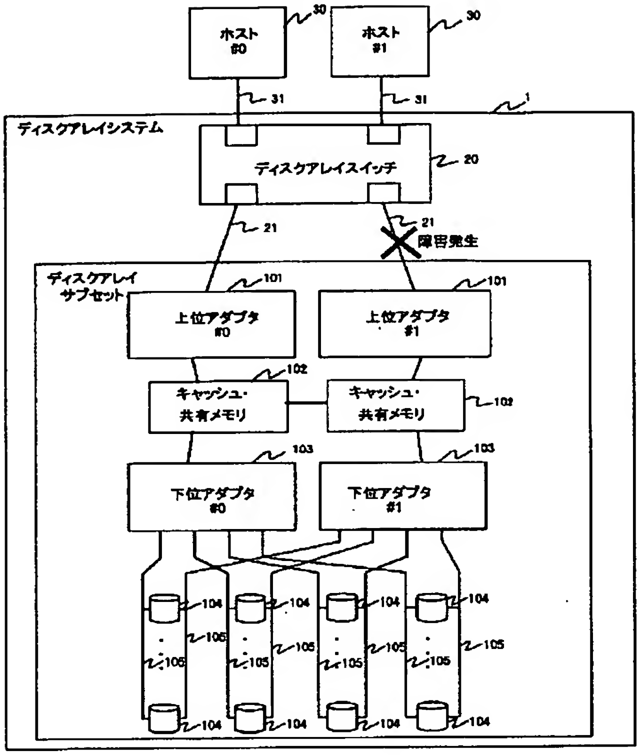
【図２５】

図25



【図２６】

図26



フロントページの続き

(72)発明者　山本　彰
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(72)発明者　味松　康行
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(72)発明者　佐藤　雅彦
神奈川県小田原市国府津2880番地　株式会社日立製作所ストレージシステム事業部内